Introduction to Statistics for Artificial Intelligence

Yonghyun Kwon · Sungwan Bang · Rosy Oh · Jaeshin Park

First Edition

Published by Kyungmoon.

ISBN: 979-11-6073-843-8 *First Edition, 2026.*

Contents

Preface

Statistics is essential for understanding data, uncertainty, and decision-making. In the era of artificial intelligence, its importance is even greater, because AI systems rely on statistical ideas such as probability, estimation, prediction, and inference. This book aims to introduce these concepts clearly and practically for students beginning their study of statistics and AI.

In preparing this textbook, we referred to *OpenIntro Statistics* by Diez, Barr, and Çetinkaya-Rundel (2019), which provided valuable guidance in presenting introductory statistical concepts in an accessible and systematic way.

We would like to express our sincere gratitude to the professors of the Department of Mathematics at the Korea Military Academy (KMA), as well as to the cadets of KMA. Their dedication to teaching and learning has provided the motivation and academic environment for the development of this book.

Most of the exercise problems in this book are based on problems used in past examinations. This book was developed with the assistance of artificial intelligence, which supported the drafting, organization, and refinement of the material. We hope it will serve as a useful foundation for further study in statistics, artificial intelligence, and data-driven decision-making.

How to use this book. Each chapter interleaves several kinds of material. Core ideas are stated as boxed *definitions* and *key results*; *examples* and *guided practice* work through them step by step; and every section ends with *exercises* whose full worked solutions are collected in Appendix C. The statistical software used throughout this book is **R**, a free and open-source language for statistical computing and graphics (`https://www.r-project.org`). Short blocks of *R output* show how the same computations are carried out in practice. A brief, self-contained introduction to **R**—enough to reproduce every computation in the text—is given in Appendix A. You will learn the most by attempting the guided practice and exercises on your own before turning to the solutions.

CHAPTER 1

Introduction to Data

Statistics is the science of learning from data, and it is the foundation on which modern artificial intelligence is built. Before we can analyze, model, or draw conclusions, we must understand what data *are*, how they are structured, and where they come from. This chapter builds the vocabulary of statistics from the ground up: data and the variables that describe them, the distinction between populations and samples, and the critical difference between studies that can establish *causation* and those that can only reveal *association*. By the end, you will be able to classify variables, identify the scope of a study, and reason carefully about what conclusions the data can — and cannot — support.

1.1 Statistics and artificial intelligence

Every day we face questions that cannot be answered by examining a single number or a single event. Does a new training program actually improve performance? Which production line is responsible for an elevated defect rate? **Statistics** is the mathematical and methodological framework for collecting, organizing, analyzing, and drawing conclusions from data in the presence of uncertainty.

This same framework is the foundation of **artificial intelligence** (AI). A modern AI system is not programmed with hand-written rules; it *learns* from data. Training a model means estimating its

unknown parameters from observed examples, and using the model means predicting the outcome for new, unseen cases — the very logic by which statistics uses a *sample* to reason about a *population*. Probability measures the uncertainty in those predictions, estimation selects the values that best fit the data, and inference tells us how far the conclusions can be trusted. The central ideas of this book — probability, estimation, prediction, and inference — are therefore exactly the ideas on which machine learning rests.

Within statistics itself, two broad branches address different stages of this process.

DEFINITION 1.1 • Descriptive and Inferential Statistics

- **Descriptive statistics** summarizes and describes the main features of a dataset using numerical measures (means, proportions, correlations) and visual displays (histograms, scatter plots). It makes no claim beyond the data at hand.
- **Inferential statistics** uses information from a *sample* to draw conclusions — called **inferences** — about a larger *population*, always accompanied by a quantified measure of uncertainty such as a confidence interval or a p-value.

EXAMPLE 1.1 • Two Branches in Practice

A researcher records the resting heart rates of 80 cadets and computes the sample mean of 68 beats per minute. Reporting that mean is **descriptive**: it summarizes what was observed. Using that sample mean to conclude "the average resting heart rate of *all* cadets is close to 68 bpm, with a margin of error of ±2 bpm" is **inferential**: the claim extends beyond the 80 individuals who were measured.

REMARK 1.1 • Why both branches matter

Descriptive statistics is the essential first step: before making inferences — or training a model — always explore and summarize the data. Skipping this step can hide data-entry errors, outliers, or unexpected patterns that would invalidate later analysis.

EXERCISES (§1.1)

1. Explain the difference between *descriptive* and *inferential* statistics. For a dataset of the final-exam scores of one class, give one example of a descriptive summary and one example of an inferential claim.
2. A commander wants the average time all cadets in a large regiment take to complete an obstacle course, but can only time a sample of 50 cadets.
   (a) What is the population and what is the sample?
   (b) Is computing the average of the 50 measured times a descriptive or an inferential act?
   (c) Is using that average to estimate the regiment-wide average descriptive or inferential?

1.2 Data and variables

DEFINITION 1.2 • Data and Observations

Data are recorded values collected on **cases** (also called **observations** or **subjects**). Each case represents one individual, item, or unit in the study.
A **data matrix** is the standard rectangular layout in which

- each **row** corresponds to one case, and
- each **column** corresponds to one **variable** — a characteristic measured on every case.

EXAMPLE 1.2 • Classroom Survey

A survey of 86 students in an introductory statistics course recorded the following variables: `gender`, `sleep` (hours per night), `bedtime` (time interval), `countries` (number visited), and `dread` (self-reported dread of class, scale 1–5).

A portion of the resulting data matrix is shown below.

Student	gender	sleep	bedtime	countries	dread
1	male	5	12–2am	13	3
2	female	7	10pm–12am	7	2
3	female	5.5	12–2am	1	4
4	female	7	12–2am	0	2
⋮	⋮	⋮	⋮	⋮	⋮
86	male	6	12–2am	3	3

Row 3 is a single *observation*: one female student who sleeps 5.5 hours, goes to bed between midnight and 2 am, has visited 1 country, and rates her dread at 4 out of 5.

Not all variables behave the same way, and the type of variable determines which summaries and analyses are appropriate.

DEFINITION 1.3 • Numerical and Categorical Variables

- A **numerical** (or **quantitative**) variable takes number values for which arithmetic operations — addition, averaging — are meaningful. Numerical variables are further divided:
  - **Continuous**: can take any value in an interval; limited only by measurement precision (e.g. height, temperature, sleep duration).
  - **Discrete**: takes only countable, separated values, typically non-negative integers

(e.g. number of countries visited, defect count).

- A **categorical** (or **qualitative**) variable places each case into one of a fixed set of groups or levels. Categorical variables are further divided:
  - **Nominal**: levels have no inherent ordering (e.g. gender, blood type, branch of service).
  - **Ordinal**: levels have a natural order but the distances between levels are not necessarily equal (e.g. dread on a 1–5 scale, education level).

EXAMPLE 1.3 • Classifying the Survey Variables

Returning to the classroom survey from Example 1.2:

- `gender` — **categorical** (male/female/other; no natural ordering).
- `sleep` — **numerical, continuous** (hours can take non-integer values such as 5.5).
- `bedtime` — **categorical, ordinal** (time intervals such as 10pm–12am, 12–2am have a natural ordering, but arithmetic on intervals is not meaningful).
- `countries` — **numerical, discrete** (a whole-number count; you cannot visit 2.3 countries).
- `dread` — **categorical, ordinal** (an integer scale 1–5 with ordered levels but unequal perceived gaps between them).

REMARK 1.2 • Telephone area codes

A telephone area code looks numerical — it is written as a three-digit number — but arithmetic on it is meaningless: the average of area codes 02 and 051 is not a useful quantity, and a larger code does not indicate "more" of anything. Area codes are therefore **categorical**, not numerical. The general rule: ask whether addition and averaging make substantive sense. If not, the variable is categorical.

VARIABLE CLASSIFICATION SUMMARY

Type	Sub-type	Key feature
Numerical	Continuous	Any value in a range
	Discrete	Countable, separated values
Categorical	Nominal	Unordered groups
	Ordinal	Ordered groups, unequal gaps

EXERCISES (§1.2)

1. A spreadsheet records, for each of 120 servers, its region, its number of CPU cores, and its average daily uptime (percent).
   (a) In the corresponding data matrix, how many rows and how many columns are there?
   (b) What does a single row represent?
   (c) Classify each of the three variables as numerical or categorical.
2. Classify each variable as numerical (continuous or discrete) or categorical (nominal or ordinal), justifying each answer in one sentence.
   (a) The fuel consumption of a vehicle (litres per 100 km).
   (b) The number of textbooks a student owns.
   (c) A soldier's rank (Private, Corporal, Sergeant, . . .).
   (d) The serial number printed on a piece of equipment.
   (e) The time (in minutes) to complete an obstacle course.

1.3 Populations, samples, and study design

Statistical questions are almost always about a group too large or too costly to examine entirely. The solution is to study a smaller, representative subset.

DEFINITION 1.4 • Population and Sample

The **population** is the complete collection of all cases about which conclusions are desired. A **sample** is a subset of the population that is actually observed.

- A **parameter** is a numerical summary computed from the full population (usually unknown).
- A **statistic** is a numerical summary computed from the sample (observable; used to estimate the parameter).

EXAMPLE 1.4 • Estimating Cadet Height

Suppose we wish to estimate the average height of all cadets in a military academy. Measuring every cadet is feasible but time-consuming; instead we measure a randomly chosen group of 50 cadets.

- **Population:** all cadets in a military academy.
- **Sample:** the 50 cadets who were measured.
- **Parameter:** the true mean height μ of all cadets (unknown).
- **Statistic:** the sample mean height $\bar{x}$ computed from the 50 measurements (our estimate of μ).

REMARK 1.3 • Scope of inference

Inferences are valid only for the population from which the sample was drawn. If a study samples only adult women runners, its conclusions about running efficiency cannot automatically be generalized to men, children, or non-runners. Matching the target population to the sampled population is a fundamental responsibility of study design.

A good sample tells us *what* is true of a population; a further question is *why* two characteristics move together.

DEFINITION 1.5 • Association and Causation

Two variables are **associated** (or **dependent**) when the value of one tends to change predictably with the value of the other; if no such relationship is apparent, they are **independent**. An observed association between X and Y is consistent with three explanations: X causes Y, Y causes X, or a **confounding variable** Z influences both and creates a spurious link. Association alone therefore **never establishes causation**.

DEFINITION 1.6 • Observational Study and Experiment

In an **observational study** the researcher measures variables as they naturally occur, without intervening; confounding variables generally cannot be ruled out, so causal conclusions require additional, often untestable, assumptions. In a **randomized experiment** the researcher assigns cases to treatment conditions using **randomization**; because randomization balances confounders across groups on average, a significant difference between groups can — under a valid design — be attributed to the treatment. Randomized experiments thus give the strongest basis for a causal claim; observational and quasi-experimental designs can support one only under explicit, carefully argued assumptions.

For example, weekly study hours and GPA are positively associated, yet studying more may not by itself raise GPA — an unmeasured factor such as motivation could drive both.

EXERCISES (§1.3)

1. Distinguish a *parameter* from a *statistic*.
   (a) If the regiment's true average obstacle-course time is $\mu = 14.2$ minutes and a sample of 50 cadets averages $\bar{x} = 13.8$ minutes, which value is the parameter and which is the statistic?
   (b) Explain why we usually study a sample rather than the whole population.
   (c) Give one reason a sample might fail to represent its population.
2. For each scenario, state whether the study is *observational* or *experimental* and whether a causal conclusion is warranted.
   (a) Researchers examine hospital records and find that patients who received a certain medication had shorter stays on average.
   (b) Volunteers are randomly assigned to receive a new vaccine or a placebo, and infection rates are compared after three months.

(c) A journalist notes that cities with more coffee shops tend to have higher average incomes.

CHAPTER 2

Summarizing Data

Once data have been collected, the first job of a statistician is not to test hypotheses or fit models — it is simply to *look*. A raw table of numbers hides almost everything that matters: where the values cluster, how far they spread, whether the data are symmetric or lopsided, and whether a few unusual cases dominate the picture. This chapter develops the standard toolkit for **exploratory data analysis**: numerical summaries of *center* and *spread*, graphical displays for one and two variables, and the vocabulary of *distribution shape*. We treat numerical and categorical variables in turn, and throughout we lean on two running examples drawn from machine learning — the response times of a deployed model and the predictions of an image classifier.

2.1 Examining numerical data

Throughout this section we study a single numerical variable: the **inference latency** of a deployed neural network, measured in milliseconds (ms), for each of $n = 500$ requests sent to a model-serving API. (The data are simulated but reflect the behavior of a real service.) Latency matters in practice — a recommendation model that takes too long to respond degrades the user experience — so an engineer naturally asks: what is a *typical* response time, how *variable* is it, and how often is a request

alarmingly slow?

2.1.1 Dot plots, histograms, and the shape of a distribution

The simplest picture of a numerical variable is a **dot plot**, which places one dot above the number line for each observation, stacking dots when values coincide. Figure 2.1 shows a small sample of 20 latencies.

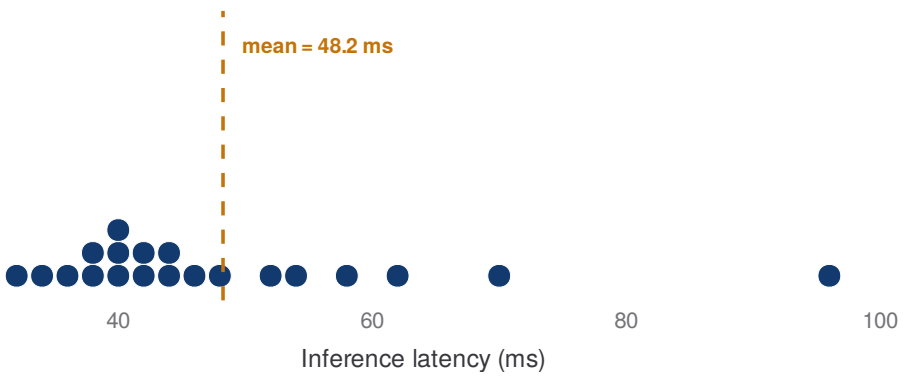


Figure 2.1. Dot plot of 20 inference latencies. Each dot is one request; the dashed line marks the mean. The single slow request near 96 ms pulls the mean to the right of the bulk of the data.

The most familiar summary of center is the **mean**.

DEFINITION 2.1 • Sample mean

For observations $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

We reserve the Greek letter μ for the (usually unknown) **population mean**; $\bar{x}$ is its sample-based estimate.

For the 20 latencies in Figure 2.1, the values sum to 965 ms, so $\bar{x} = 965/20 = 48.25 \approx 48.2$ ms. Mechanically, the mean is the *balance point* of the dot plot: if the dots were equal weights resting on a ruler, the ruler would balance at $\bar{x}$. This is exactly why a single heavy value far to the right (the 96 ms request) tips the balance point away from where most of the data sit — an observation we will return to when we discuss robustness.

REMARK 2.1 • Stacked dot plots and larger samples

Dot plots are ideal for small or moderate samples. With hundreds or thousands of observations the dots merge into a solid band and a **histogram** (next) becomes the better display.

A **histogram** groups the range of a numerical variable into adjacent intervals called **bins** and draws a bar over each bin whose height is the number of observations (the **frequency**) falling in it. Histograms reveal the overall **distribution** — the pattern of which values are common and which are rare.

The one choice a histogram forces on us is the **bin width**. Figure 2.2 shows the full set of 500 latencies at three different bin widths.

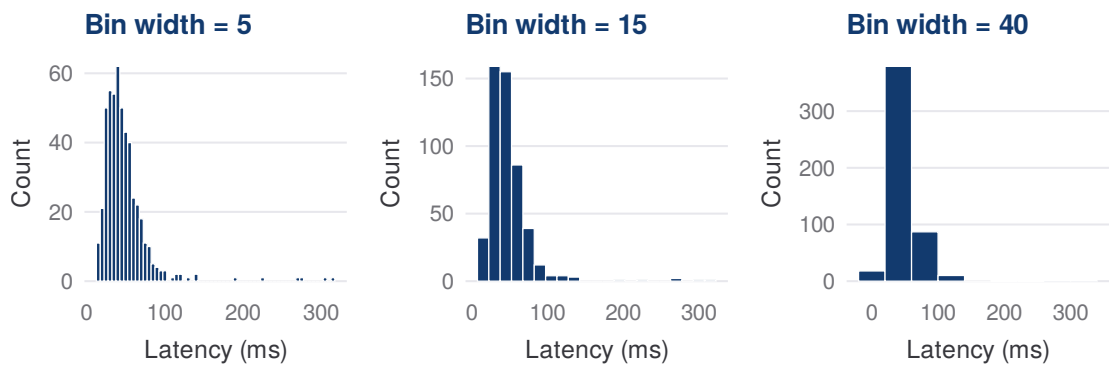


Figure 2.2. The same 500 latencies at three bin widths. Too narrow (left) is noisy; too wide (right) hides structure; an intermediate width (center) shows the shape clearly.

> **! CAUTION**
>
> There is no single "correct" bin width. Very narrow bins produce a jagged display dominated by sampling noise; very wide bins smooth away real features such as a second peak. Always try a few widths before drawing conclusions.

Whatever the bin width, the histogram lets us describe the **shape** of the distribution. Three aspects of shape are worth naming.

Modality. A **mode** is a prominent peak. A distribution is **unimodal** with one peak, **bimodal** with two, and **multimodal** with several; a distribution with no peak — roughly flat — is called **uniform**. Figure 2.3 illustrates each. Multiple modes often signal that the data mix together two or more distinct subpopulations (for instance, requests served from a warm cache versus a cold start).

Skewness. A distribution is **symmetric** if its left and right halves are approximate mirror images. It is **right-skewed** (positively skewed) if it has a long tail stretching to the right, and **left-skewed** if the long tail points left (Figure 2.4). Latency data are almost always right-skewed: most requests are fast, but a few are very slow.

Unusual observations. Values that fall far from the bulk of the data are called **outliers**. Outliers may be genuine (a real, exceptionally slow request) or erroneous (a logging glitch). Either way they deserve attention: they can distort summaries and they are sometimes the most interesting cases in the dataset — in fraud detection or fault monitoring, the outliers are the whole point.

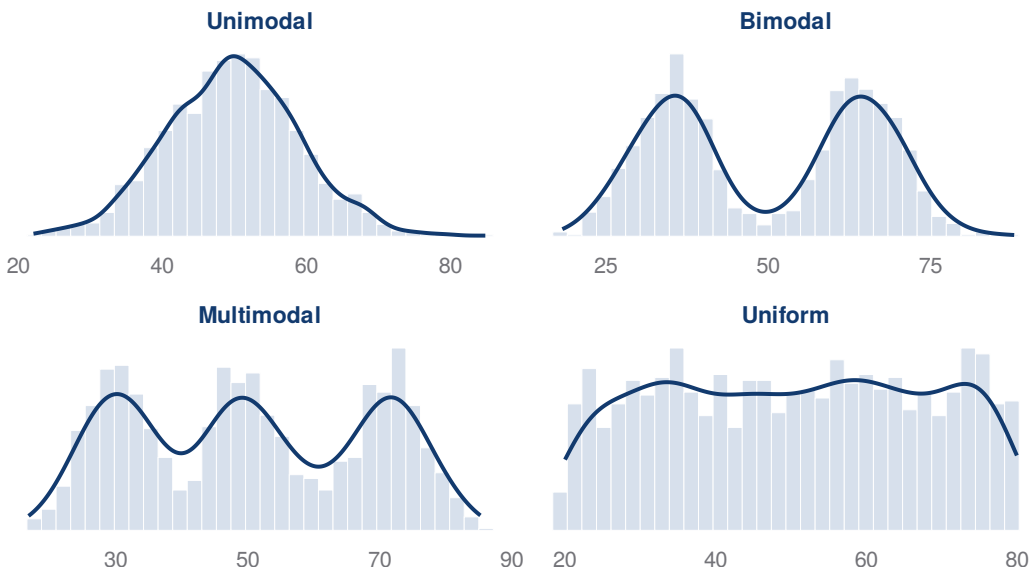


Figure 2.3. Four modality patterns. Multiple peaks frequently indicate a mixture of distinct groups within the data.

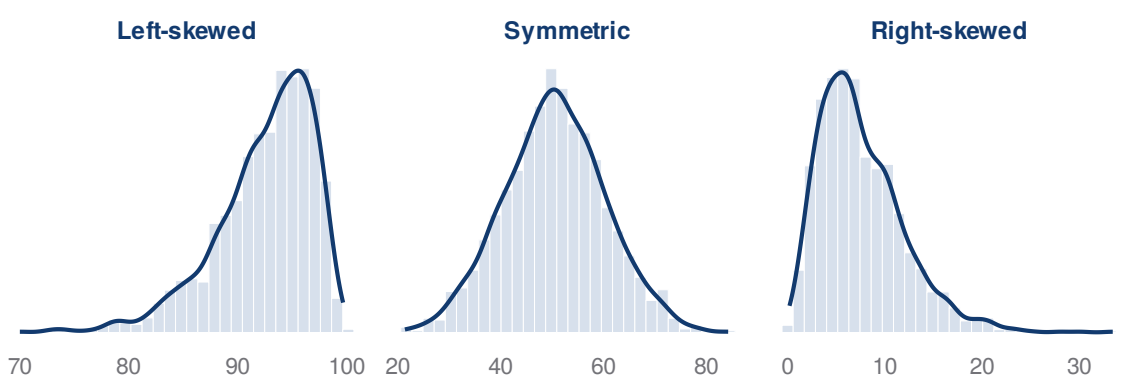


Figure 2.4. Left-skewed, symmetric, and right-skewed distributions. The skew is named for the direction of the *long tail*.

GUIDED PRACTICE 2.1

Sketch what you would expect the distribution of the following variables to look like, and name the likely shape:

(a) The file sizes (in kB) of images uploaded to a website.

(b) The scores of a large class on an easy exam (out of 100).

(c) Human body temperatures (°C) of healthy adults.

Answers: (a) right-skewed — many small files, a few very large ones; (b) left-skewed — most scores near the top, a tail of low scores; (c) roughly symmetric and unimodal about 37 °C.

2.1.2 Variance and standard deviation

Center alone is not enough: two services could share a mean latency of 48 ms while one is metronome-steady and the other swings wildly. **Spread** (or **variability**) measures how far the data fall from their center. The dominant measures of spread are built from the **deviations** $x_i - \bar{x}$.

DEFINITION 2.2 • Sample variance and standard deviation

The **sample variance** is the (almost-)average squared deviation,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

and the **sample standard deviation** is its square root, $s = \sqrt{s^2}$. We write σ^2 and σ for the corresponding population quantities.

Squaring the deviations makes every contribution positive (so that values above and below the mean do not cancel) and penalizes large deviations heavily. The standard deviation s then returns us to the original units. For our 500 latencies, $s \approx 31.2$ ms; loosely, a typical request differs from the mean of 48.4 ms by about 31 ms.

READING THE STANDARD DEVIATION

- s is in the *same units* as the data (s^2 is in squared units).
- $s \geq 0$, and $s = 0$ only if every value is identical.
- For roughly symmetric, bell-shaped data, about two-thirds of observations lie within one standard deviation of the mean and about 95% within two. (This rule of thumb fails badly for strongly skewed data such as latency — a reason to be cautious.)

REMARK 2.2 • Why divide by $n-1$?

Dividing by $n-1$ rather than n corrects a subtle downward bias: the deviations are taken from $\bar{x}$ (computed from the same data) rather than from the true mean μ, which makes the squared deviations slightly too small on average. The factor $n-1$, called the **degrees of freedom**, compensates. We return to this point when we study estimation.

2.1.3 The median, quartiles, and robust statistics

Because squared deviations magnify outliers, the mean and standard deviation can be misleading for skewed data. A second family of summaries is based on *order* rather than arithmetic.

DEFINITION 2.3 • Median and quartiles

Sort the data from smallest to largest.

- The **median** (Q_2) is the middle value: the average of the two middle values when n is even.
- The **first quartile** Q_1 is the median of the lower half; the **third quartile** Q_3 is the median of the upper half.
- The **interquartile range** is $\mathrm{IQR} = Q_3 - Q_1$, the spread of the middle 50% of the data.

Together, (min, Q_1, Q_2, Q_3, max) form the **five-number summary**.

For the 500 latencies the five-number summary (in ms) is approximately

$$\text{min} = 14, \quad Q_1 = 31.5, \quad \text{median} = 42.3, \quad Q_3 = 56.0, \quad \text{max} = 318,$$

so $\mathrm{IQR} = 56.0 - 31.5 = 24.5$ ms. The five-number summary is displayed compactly by a **box plot**.

DEFINITION 2.4 • Box plot

A **box plot** draws a box from Q_1 to Q_3 with a line at the median. **Whiskers** extend from the box to the most extreme observation within $1.5 \times \mathrm{IQR}$ of the nearer quartile; any observation beyond a whisker is plotted individually as a suspected **outlier**.

The box plot makes the right skew unmistakable: the median sits left of center within the box, the upper whisker is longer than the lower, and a string of outliers trails off to the right. A box plot discards fine detail (it cannot show modality, for example) but it is unmatched for comparing several groups at a glance, as we will see in Section 2.2.

A summary is **robust** if it changes little when a small number of observations take extreme values. The median and IQR are robust; the mean and standard deviation are not.

Figure 2.6 overlays the mean and median on the latency histogram. The mean (48.4 ms) is dragged noticeably to the right of the median (42.3 ms) by the slow-request tail, even though those slow

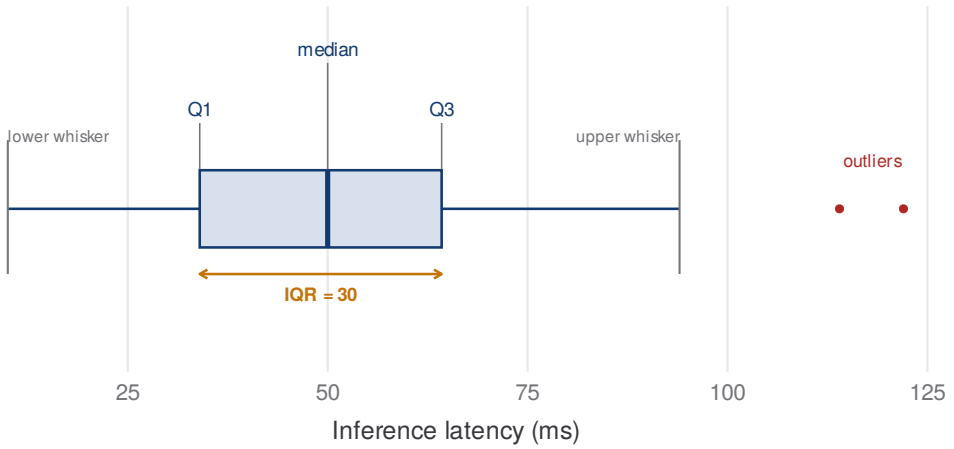


Figure 2.5. Anatomy of a box plot for a different group of latencies ($Q_1 = 35$, $Q_3 = 65$, $\text{IQR} = 65-35 = 30$). The box spans the interquartile range; whiskers reach the farthest non-outlying points ($Q_1 - 1.5\times\text{IQR} = -10$, $Q_3 + 1.5\times\text{IQR} = 110$); individual outliers (red) are the slow requests in the right tail.

requests are rare.

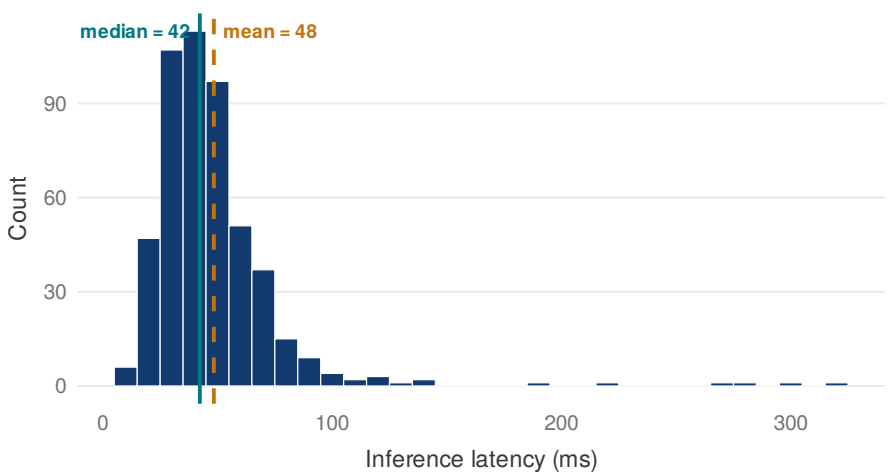


Figure 2.6. For right-skewed data the mean (amber, dashed) is pulled toward the long tail and exceeds the median (teal, solid). The median better represents a "typical" value here.

WHICH SUMMARY SHOULD I REPORT?

	Center	Spread
Sensitive to outliers	mean $\bar{x}$	standard deviation s
Robust to outliers	median Q_2	IQR

For roughly symmetric data, report the mean and standard deviation. For skewed data or data with outliers, prefer the median and IQR — or report both and explain the difference.

EXAMPLE 2.1 • One slow request

Suppose a single new request is logged at a catastrophic 5,000 ms (a timeout). Adding it to the 500 latencies barely moves the median — the middle of an ordered list shifts by at most one position — but it raises the mean by roughly 10 ms and inflates the standard deviation substantially. A monitoring dashboard that tracked only the mean would sound an alarm; one that tracked the median would correctly report that typical performance was unchanged. Both views are useful, but only if their different sensitivities are understood.

2.1.4 Transformations and scatterplots

Strong skew makes data hard to visualize and violates the symmetry assumptions of many later methods. A **transformation** re-expresses each value on a new scale to tame the skew. The most common choice for positive, right-skewed data is the **logarithm**.

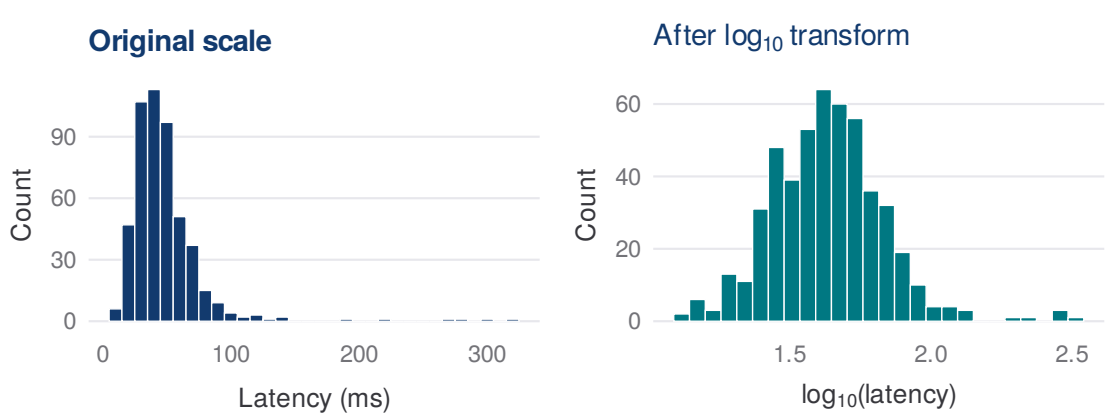


Figure 2.7. The right-skewed latencies (left) become approximately symmetric after a base-10 log transform (right). Equal multiplicative steps — ×10 — become equal additive steps on the log scale.

A log transform compresses the long right tail and stretches the crowded left side, often producing the near-symmetric, bell-shaped distribution that downstream analyses prefer. Log scales are pervasive in machine learning: learning rates, model sizes, and loss values are routinely plotted on logarithmic axes for exactly this reason.

> **! CAUTION**
>
> A transformation changes the scale, so summaries must be interpreted on that scale. The mean of the logs is *not* the log of the mean. When reporting results, either state clearly that values are on the log scale or transform back to the original units.

So far we have summarized a single numerical variable. When two numerical variables are measured on each case, the natural display is a **scatterplot**, which plots one variable on each axis and one point per case. Figure 2.8 shows two measured features — petal length and petal width — for 150 iris flowers, a classic dataset in pattern recognition.

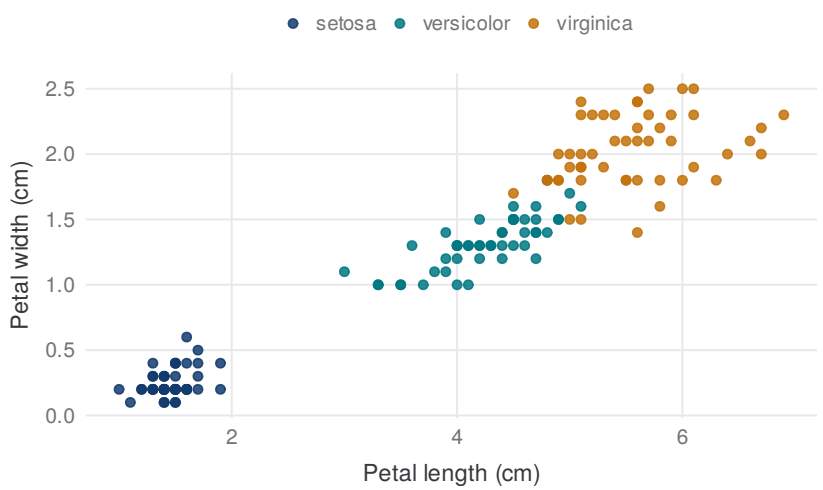


Figure 2.8. Two features of 150 iris flowers. Petal length and width are strongly positively associated, and the three species form visibly separated clusters — the property that makes this a textbook classification problem.

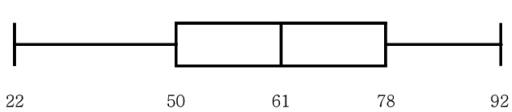
Petal length and width rise together: the variables are **positively associated**. A scatterplot can also reveal the *form* of a relationship (linear or curved), its *strength* (tight or diffuse), and any outliers or clusters. Here the points separate into three clouds, one per species — a preview of why this dataset is a staple benchmark for classification algorithms. We quantify such relationships with **correlation** and **regression** in a later chapter.

EXERCISES (§2.1)

1. The table below shows descriptive statistics for chick weights (in grams) when soybean meal was used as a feed additive. Based on these values, draw a boxplot for chick weights.

Minimum	Q_1	Median (Q_2)	Q_3	Maximum	Mean
158	206.75	248.0	270.00	329	246.43

2. A group of 100 soldiers in a training company did a sit-up test. The results are shown in the boxplot below. Based on the grading standards in the table, fill in the blanks.



Grade	Fail	Grade 3	Grade 2	Grade 1	Special
Counts	61 or fewer	62–69	70–77	78–85	86 or more

(a) The first quartile (Q_1) of the number of sit-ups is ( ).

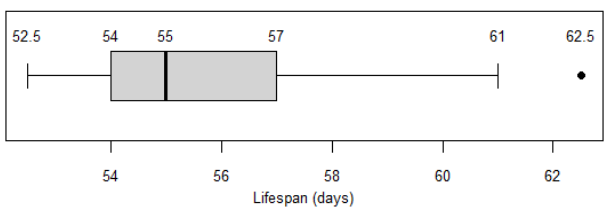
(b) The interquartile range (IQR) of the number of sit-ups is ( ).

(c) The number of soldiers in the company who are Grade 1 or Special grade is ( ).

(d) The fewest sit-ups performed by any soldier in the company is ( ).

(e) The number of soldiers in the company who failed the sit-up test is ( ).

3. A boxplot of aircraft tire lifespans produced by Company A is given below. Determine whether each of the following statements is true (O) or false (X).



(a) The median lifespan of aircraft tires produced by Company A is 55 days. (O / X)

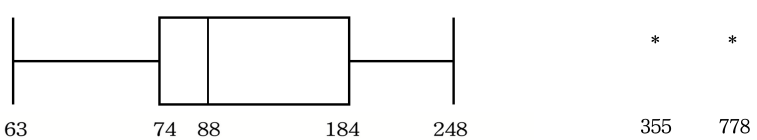
(b) The interquartile range (IQR) of the lifespan is 3 days. (O / X)

(c) The minimum lifespan is 54 days. (O / X)

(d) Approximately 25% of the tires have a lifespan shorter than 55 days. (O / X)

(e) The distribution of the lifespan is left-skewed. (O / X)

4. The following boxplot shows the population data (unit: thousand people) of 50 cities in a country in 1960. Fill in the blanks with the appropriate values.



(a) The median population is ( ) thousand people.

(b) The interquartile range (IQR) of the population is ( ) thousand people.

(c) The upper whisker limit ($Q_3 + 1.5 \times \text{IQR}$) is ( ) thousand people.

(d) The number of outliers is ( ).

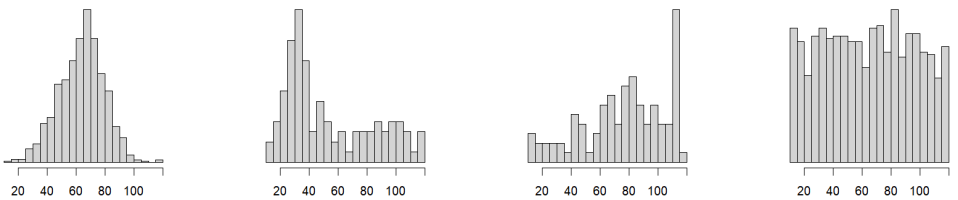
(e) The difference between the maximum population and the minimum population is ( ) thousand people.

5. The daily air quality index (AQI) was reported for a sample of 91 days in 2024 in a certain city. Summary statistics of the daily AQI are provided below.

Minimum	Q_1	Median	Mean	Q_3	Maximum
10	30	40	50	80	120

(a) Draw a boxplot for the daily AQI of this sample. (There is no outlier in the sample.)

(b) Identify which of the following corresponds to the histogram of the daily AQI in the sample.



(A) (B) (C) (D)

2.2 Considering categorical data

We now turn to **categorical** variables, whose values are labels rather than numbers. Our running example is the evaluation of an image classifier on a labeled test set of 600 images belonging to three classes — *cat*, *dog*, and *bird*. Each image has two categorical attributes: its *true class* (the correct label) and the model's *predicted class*. Comparing them is the central task of evaluating any classifier.

2.2.1 Contingency tables and proportions

A **contingency table** (or **two-way table**) tabulates the number of cases for every combination of two categorical variables. When the two variables are the true and predicted labels of a classifier, the contingency table is exactly what machine-learning practitioners call a **confusion matrix**.

Table 2.1. Confusion matrix of the image classifier: counts of the 600 test images by true class (rows) and predicted class (columns), with row and column totals.

	Predicted			
True	Cat	Dog	Bird	**Total**
Cat	182	12	6	200
Dog	15	210	25	250
Bird	9	14	127	150
Total	206	236	158	600

The row and column totals are the **marginal** counts; each gives the distribution of one variable on its own (for example, the test set contains 200 cats, 250 dogs, and 150 birds). The interior cells are **joint** counts. Dividing counts by a total turns them into proportions, and *which* total we divide by answers a different question.

DEFINITION 2.5 • Joint, marginal, and conditional proportions

In a contingency table:

- a **joint proportion** divides a cell by the grand total;

- a **marginal proportion** divides a row or column total by the grand total;
- a **conditional proportion** divides a cell by its own *row* total (a **row proportion**) or its own *column* total (a **column proportion**).

The diagonal cells are the correct predictions. Their joint proportion is the model's **accuracy**,

$$\text{accuracy} = \frac{182 + 210 + 127}{600} = \frac{519}{600} = 0.865,$$

so the classifier labels 86.5% of images correctly. But accuracy hides where the errors lie, and the two kinds of conditional proportion expose them:

- A **row proportion** answers "of all true cats, what fraction were labeled cat?" — the **recall** for that class. For cats it is $182/200 = 0.910$.
- A **column proportion** answers "of all images called cat, what fraction really were cats?" — the **precision** for that class. For cats it is $182/206 = 0.883$.

CHOOSING THE APPROPRIATE PROPORTION

The right conditioning depends on the question. *Recall* (row proportion) measures how well the model *finds* a class; *precision* (column proportion) measures how much to *trust* a positive prediction. They generally differ. Here the bird class has high recall ($127/150 = 0.847$) but lower precision ($127/158 = 0.804$): the model rarely misses a bird, but a fair number of its "bird" calls are actually dogs.

! CAUTION

Never compare raw counts across groups of different sizes. Because the test set has more dogs (250) than birds (150), the dog class will have more errors *in absolute terms* almost automatically. Convert to the appropriate conditional proportion before comparing.

2.2.2 Bar plots and pie charts

For a *single* categorical variable, a **bar plot** draws one bar per category with height equal to its count (or proportion). Figure 2.9 shows the class distribution of the test set.

A **pie chart** displays the same information as wedges of a circle (Figure 2.10). Pie charts are popular but hard to read: the eye judges lengths along a common scale far more accurately than it judges angles or areas.

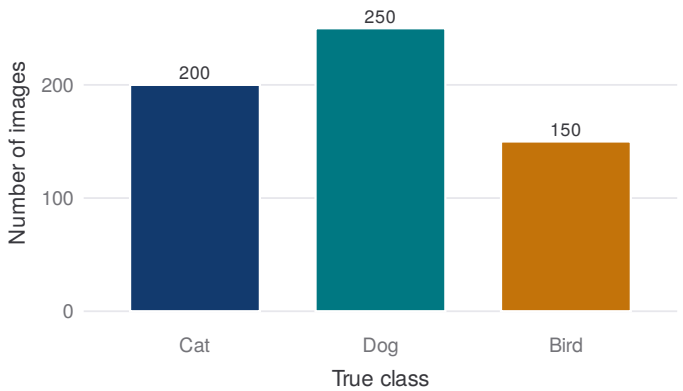


Figure 2.9. Bar plot of the test set's true-class distribution. Bars share a common baseline, so heights are easy to compare.

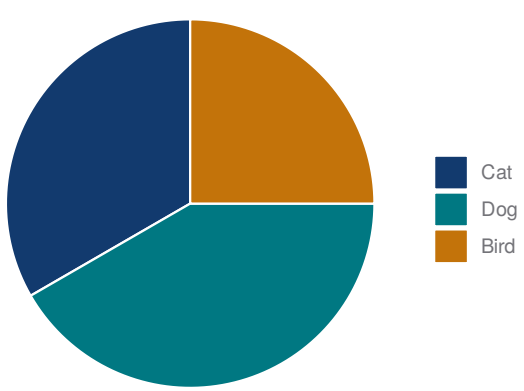


Figure 2.10. Pie chart of the same class distribution. Comparing the dog and cat wedges is harder than comparing the corresponding bars.

> **REMARK 2.3 • Prefer bars to pies**
>
> Reserve pie charts for the rare case of a few categories that sum to a meaningful whole and differ greatly in size. In almost every other situation a bar plot communicates more clearly.

To display two categorical variables together we extend the bar plot. A **grouped** (side-by-side) bar plot places the bars for one variable next to each other within each level of the other (Figure 2.11); it is best for comparing raw counts.

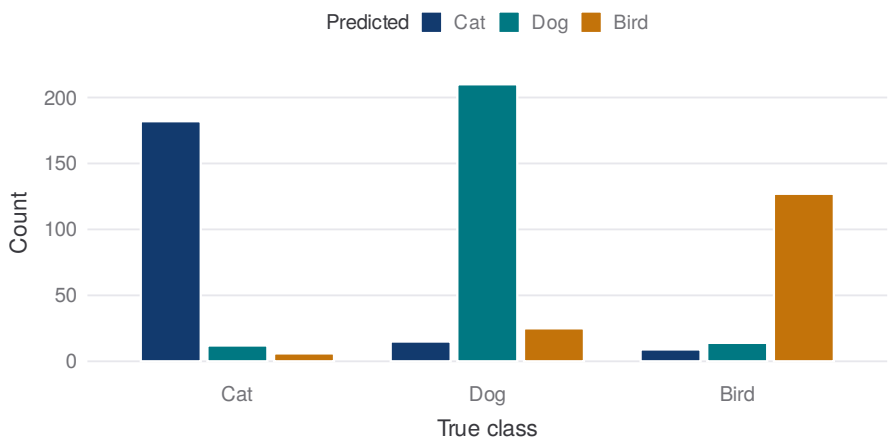


Figure 2.11. Grouped bar plot of predicted class within each true class. The tall on-diagonal bars are correct predictions; the short bars are confusions.

A **standardized stacked** bar plot instead stacks the bars to a common height of 100%, so each bar shows the *conditional* (row) proportions (Figure 2.12). This is the natural display when the question is about proportions rather than counts — here, each bar is precisely the recall breakdown for one true class.

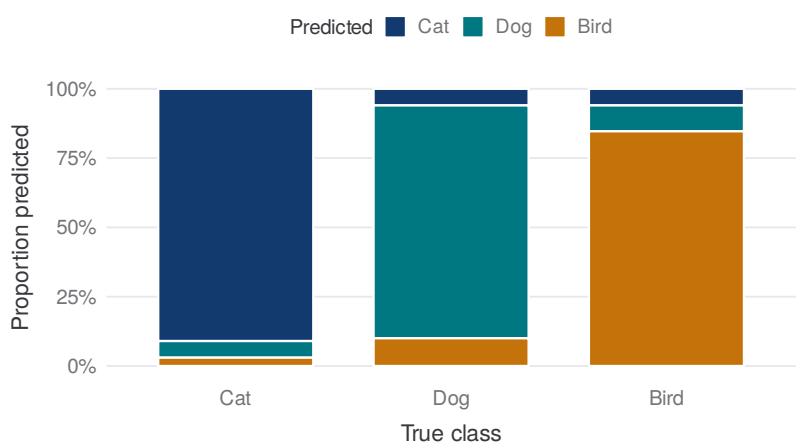


Figure 2.12. Standardized stacked bar plot. Each bar sums to 100% and shows, for one true class, the proportions predicted into each class.

COUNTS OR PROPORTIONS?

Use a grouped bar plot to compare counts and a standardized stacked bar plot to compare proportions. If group sizes differ, the standardized version is usually the more honest comparison.

2.2.3 Comparing a numerical variable across groups

Finally, categorical and numerical data meet when we ask how a numerical variable differs across the levels of a categorical one. The standard tool is a set of **side-by-side box plots**, one per group on a shared axis. Figure 2.13 compares petal length across the three iris species.

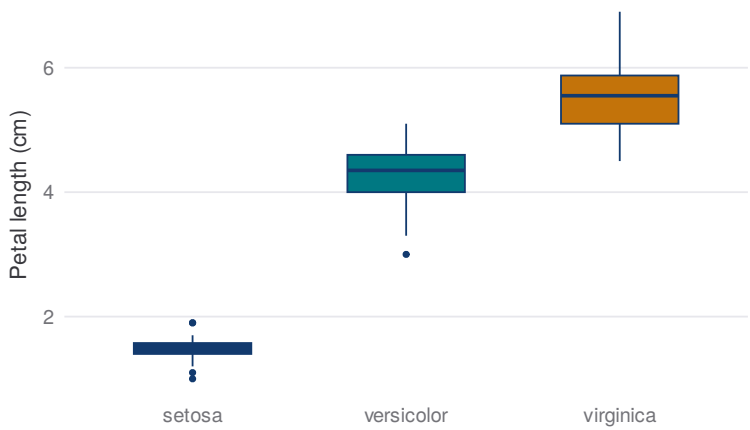


Figure 2.13. Side-by-side box plots of petal length by iris species on a shared axis. The near-total separation of *setosa* from the others is what lets a classifier identify it almost perfectly.

The three boxes barely overlap, which tells us at a glance that petal length is a highly *discriminative* feature: knowing it goes a long way toward identifying the species. Side-by-side box plots are therefore a quick, model-free way to judge which features are likely to be useful before any algorithm is trained.

EXERCISES (§2.2)

1. Using Table 2.1, compute the recall and precision for the *dog* class. State in words what each number means.
2. Of the 600 images, what proportion are *both* truly a bird *and* predicted to be a bird? Is this a joint, marginal, or conditional proportion?

CHAPTER 3

Probability

Probability is the mathematical language of uncertainty. In this chapter we build the foundations from the ground up: sample spaces, events, and the axioms that govern how probabilities combine. We then study conditional probability and conclude with Bayes' theorem — the engine of rational belief revision. By the end, you will be able to compute and interpret probabilities for real-world experiments and reason carefully about uncertain outcomes.

3.1 Defining probability

Every day we make decisions under uncertainty. Will a component fail before the next maintenance cycle? Is the email flagged by the spam filter actually spam? **Probability** is the mathematical framework for describing, quantifying, and reasoning about that uncertainty.

Any **random experiment** produces an outcome that cannot be determined in advance, even though the set of all possible outcomes is fully known. The first step in any probability problem is to name those possible outcomes precisely.

DEFINITION 3.1 • Sample Space and Event

The **sample space** Ω of a random experiment is the exhaustive collection of all possible outcomes. An **event** A is any subset $A \subseteq \Omega$.

- A **simple event** contains exactly one outcome.
- A **compound event** contains two or more outcomes.
- The **complement** of A, written $A^c = \Omega \setminus A$, is the event that A does *not* occur.

EXAMPLE 3.1 • Two-Factor Authentication

A security system runs two independent checks: a password check (Pass/Fail) and a one-time-code check (Pass/Fail). Outcomes are recorded as ordered pairs (password result, code result).

Sample space: $\Omega = \{(P,P), (P,F), (F,P), (F,F)\}$.

Events of interest:

- A = "both checks pass" = $\{(P,P)\}$ (simple event).
- B = "at least one check passes" = $\{(P,P),(P,F),(F,P)\}$ (compound event).
- A^c = "not both checks pass" = $\{(P,F),(F,P),(F,F)\}$.

Note that $A \subset B$: whenever both checks pass, at least one certainly passes.

Set operations transfer naturally to events. If A and B are events in Ω:

- **Union** $A \cup B$: event A *or* B occurs.
- **Intersection** $A \cap B$: both A *and* B occur.
- **Complement** A^c: A does *not* occur.

REMARK 3.1 • Disjoint events

Events A and B are **mutually exclusive** (or **disjoint**) if $A \cap B = \varnothing$: they cannot both occur in the same trial. In particular $\mathbb{P}(A \cap B) = 0$; the converse need not hold, since an event of probability 0 need not be empty.

Now that we have a sample space and events, we assign numbers to them. Not every assignment is legitimate — the numbers must satisfy a coherent set of rules known as the **Kolmogorov axioms**.

DEFINITION 3.2 • Probability Measure

A **probability measure** on Ω is a function $\mathbb{P}$ that assigns a real number to each event $A \subseteq \Omega$

satisfying:

(K1) Normalization: $\mathbb{P}(\Omega) = 1$.

(K2) Non-negativity: $\mathbb{P}(A) \geq 0$ for all A.

(K3) Countable additivity: if $A_1, A_2, \ldots$ are pairwise disjoint, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

The following useful results all follow directly from the axioms.

ELEMENTARY CONSEQUENCES

Let $\mathbb{P}$ be a probability measure on Ω. For any events $A, B \subseteq \Omega$:

(i) **Complement rule:** $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(ii) $\mathbb{P}(\varnothing) = 0$.

(iii) **Monotonicity:** if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(iv) **Addition rule:** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

PROOF (i) Since A and A^c are disjoint with $A \cup A^c = \Omega$, axiom (K3) gives $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$.

(iv) Write $A \cup B = A \cup (B \setminus A)$ as a disjoint union, and $B = (A \cap B) \cup (B \setminus A)$ as another. Applying (K3) to both and eliminating $\mathbb{P}(B \setminus A)$ yields the result. Parts (ii) and (iii) follow from (K3) similarly. ■

The axioms become powerful once we combine them with careful counting and the complement strategy.

EXAMPLE 3.2 • Trainee Qualification

In a group of 200 trainees, 72 qualified on an obstacle course (event A), 58 on a marksmanship test (event B), and 20 on *both*. If one trainee is selected at random, what is the probability they qualified on the obstacle course *or* the marksmanship test?

SOLUTION **Step 1.** Convert counts to probabilities:

$$\mathbb{P}(A) = \tfrac{72}{200}, \quad \mathbb{P}(B) = \tfrac{58}{200}, \quad \mathbb{P}(A \cap B) = \tfrac{20}{200}.$$

Step 2. Apply the addition rule:

$$\mathbb{P}(A \cup B) = \tfrac{72}{200} + \tfrac{58}{200} - \tfrac{20}{200} = \tfrac{110}{200} = 0.55.$$

Sanity check. $72 + 58 = 130$ counts the 20 double-qualified trainees twice; $130 - 20 = 110$

and $110/200 = 0.55$. ■

REMARK 3.2 • The "at least one" complement strategy

A recurring calculation pattern is:

$$\mathbb{P}(\text{at least one}) \;=\; 1 - \mathbb{P}(\text{none}).$$

When trials are independent, $\mathbb{P}(\text{none})$ factors into a single product, making the calculation simple even for many trials.

EXAMPLE 3.3 • At Least One Defective Item

Each flash drive is independently defective with probability 0.05. You purchase a pack of 3. What is the probability at least one is defective?

■ **SOLUTION** **Step 1.** Name the complement: "none are defective" means all three are good. **Step 2.** By independence,

$$\mathbb{P}(\text{none defective}) = (0.95)^3 \approx 0.8574.$$

Step 3. Apply the complement rule:

$$\mathbb{P}(\text{at least one defective}) = 1 - (0.95)^3 \approx 0.1426.$$

There is roughly a 14.3% chance the pack contains at least one defective drive. ■

GUIDED PRACTICE 3.1

In a class of 150 students, 54 take calculus (event C), 63 take physics (event P), and 18 take both.

(a) What is $\mathbb{P}(C \cup P)$?

(b) What is $\mathbb{P}\big((C \cup P)^c\big)$?

(c) Verify your answer to (a) using the counts directly.

Answers: (a) $\frac{54+63-18}{150} = \frac{99}{150} = 0.66$. (b) $1 - 0.66 = 0.34$. (c) 99 distinct students out of 150: $99/150 = 0.66$.

When events are *independent*, joint probabilities factor — a property exploited in reliability engineering, cryptography, and statistical modeling.

DEFINITION 3.3 • Independence

Events A and B are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

A collection $\{A_i\}$ is **mutually independent** if for every finite subcollection $\{A_{i_1}, \dots, A_{i_m}\}$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_m}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_m}).$$

! CAUTION

Independent ≠ disjoint. Disjoint events *cannot* both happen; if A occurs, B is automatically ruled out, so knowing A tells you everything about B. Two disjoint events with positive probability are therefore *never* independent. Independence means the events can coexist but neither influences the other.

In addition, **pairwise independence does not imply mutual independence**. See Exercise 3 for a classical counterexample.

PRODUCT RULE FOR INDEPENDENT EVENTS

If $A_1, \dots, A_k$ are mutually independent, then

$$\mathbb{P}(A_1 \cap \cdots \cap A_k) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_k).$$

GUIDED PRACTICE 3.2

A device must pass two independent quality checks. Check 1 passes with probability 0.98; Check 2 passes with probability 0.95.

(a) What is $\mathbb{P}(\text{both pass})$?

(b) What is $\mathbb{P}(\text{at least one fails})$?

(c) If a third independent check with pass probability 0.97 is added, what is the new $\mathbb{P}(\text{all three pass})$?

Answers: (a) $(0.98)(0.95) = 0.931$. (b) $1 - 0.931 = 0.069$. (c) $(0.98)(0.95)(0.97) \approx 0.903$.

EXERCISES (§3.1)

1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a fair die. Find the probability of (a) rolling an even number, (b) rolling a number greater than 4, and (c) rolling an even number *or* a number greater than 4.
2. Prove the addition rule $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ from the Kolmogorov axioms.

3. (Pairwise vs. mutual independence.) Let $\Omega = \{1,2,3,4\}$ with uniform probability and define $A = \{1,2\}$, $B = \{1,3\}$, $C = \{1,4\}$. Show that A, B, C are pairwise independent but *not* mutually independent.
4. A battery fails in its first month with probability 0.03, independently across batteries. What is the probability that in a pack of 5 batteries *at least one* fails in the first month? *Hint:* compute $\mathbb{P}(\text{none fail}) = (0.97)^5$ first.
5. (Inclusion–exclusion.) Show that for any three events A, B, C,

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

3.2 Conditional probability

So far all probabilities have been computed using the full sample space. Often, however, we receive *partial information* — a test result, an alarm, an observation — that should update our uncertainty. **Conditional probability** is the tool for incorporating new information in a principled way.

DEFINITION 3.4 • Conditional Probability

For events A and B with $\mathbb{P}(B) > 0$, the **conditional probability** of A given B is

$$\mathbb{P}(A \mid B) \;=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

REMARK 3.3 • Restricting the sample space

When we condition on B, we *zoom in* to the portion of Ω where B occurred, and ask: within that region, how much of it is also in A? The formula implements this by dividing the measure of $A \cap B$ by the measure of B.

GENERAL MULTIPLICATION RULE

For any events A and B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B).$$

Symmetrically, if $\mathbb{P}(A) > 0$, $\mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$.

REMARK 3.4 • Independence revisited

Events A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ (equivalently, $\mathbb{P}(B \mid A) = \mathbb{P}(B)$): knowing one event occurred does not change the probability of the other. The multiplication rule then reduces to $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, recovering Definition 3.3.

READING PROBABILITIES FROM A CONTINGENCY TABLE

When two characteristics are recorded on the same individuals, a two-way table gives all three probability types directly:

- **Marginal:** row or column total ÷ grand total.
- **Joint:** cell count ÷ grand total.
- **Conditional:** cell count ÷ *relevant row or column total*.

EXAMPLE 3.4 • Fraud-Detection Confusion Matrix

A fraud-detection model was tested on 10,000 transactions.

	True fraud	True not-fraud	Total
Predicted fraud	150	30	180
Predicted not-fraud	70	9750	9820
Total	220	9780	10000

Let F = "true fraud" and $\hat{F}$ = "predicted fraud." Find: (a) $\mathbb{P}(\hat{F})$; (b) $\mathbb{P}(F \cap \hat{F})$; (c) $\mathbb{P}(F \mid \hat{F})$ (precision); (d) $\mathbb{P}(\hat{F} \mid F)$ (recall).

■ SOLUTION

(a) **Marginal:** $\mathbb{P}(\hat{F}) = 180/10000 = 0.018$.

(b) **Joint:** $\mathbb{P}(F \cap \hat{F}) = 150/10000 = 0.015$.

(c) **Precision (condition on predicted fraud):** $\mathbb{P}(F \mid \hat{F}) = 150/180 \approx 0.833$. When the model flags a transaction, it is correct 83.3% of the time.

(d) **Recall (condition on true fraud):** $\mathbb{P}(\hat{F} \mid F) = 150/220 \approx 0.682$. The model catches 68.2% of all actual fraud.

Parts (c) and (d) involve the same cell count (150) but different denominators — the denominator defines the population we are conditioning on. ■

EXAMPLE 3.5 • Checking Independence from a Table

A survey of 100 people recorded beverage preference (Coffee/Tea) and whether they exercised the previous day.

	Exercise yes	Exercise no	Total
Coffee	32	8	40
Tea	48	12	60
Total	80	20	100

Let E = "exercised" and C = "prefers coffee." Are E and C independent?

■ **SOLUTION** Marginal: $\mathbb{P}(E) = 80/100 = 0.80$.
Conditional: $\mathbb{P}(E \mid C) = 32/40 = 0.80$.
Since $\mathbb{P}(E \mid C) = \mathbb{P}(E)$, knowing someone prefers coffee gives no information about whether they exercised. The data are consistent with independence.
Product-rule check: $\mathbb{P}(E)\mathbb{P}(C) = (0.80)(0.40) = 0.32 = 32/100 = \mathbb{P}(E \cap C)$. ■

Suppose we know $\mathbb{P}(B \mid A)$ — the probability a test is positive given disease is present — but we need $\mathbb{P}(A \mid B)$ — the probability of disease given a positive test. Bayes' theorem provides a systematic way to *invert* a conditional probability.

LAW OF TOTAL PROBABILITY

Let $A_1, \ldots, A_k$ be a partition of Ω (pairwise disjoint and exhaustive) with $\mathbb{P}(A_i) > 0$ for all i. For any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i).$$

BAYES' THEOREM

Let $A_1, \ldots, A_k$ be a partition of Ω with $\mathbb{P}(A_i) > 0$. For any event B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A_i \mid B) \;=\; \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\displaystyle\sum_{j=1}^{k} \mathbb{P}(B \mid A_j)\,\mathbb{P}(A_j)}.$$

■ **PROOF** The numerator is $\mathbb{P}(B \cap A_i)$ by the general multiplication rule. The denominator equals $\mathbb{P}(B)$ by the law of total probability. Dividing gives exactly $\mathbb{P}(A_i \mid B)$ by Definition 3.4. ■

Each quantity in Bayes' theorem has a standard name:

- $\mathbb{P}(A_i)$ — the **prior probability**: our belief about cause A_i before seeing evidence.
- $\mathbb{P}(B \mid A_i)$ — the **likelihood**: how probable is evidence B if A_i is the true cause?
- $\mathbb{P}(A_i \mid B)$ — the **posterior probability**: our updated belief about A_i after seeing B.

- $\mathbb{P}(B)$ — the **normalizing constant** (marginal likelihood), ensuring posteriors sum to 1.

EXAMPLE 3.6 • Medical Screening Test

A disease affects 1% of the population. A screening test has *sensitivity* $\mathbb{P}(T \mid D) = 0.95$ and *specificity* $\mathbb{P}(T^c \mid D^c) = 0.90$ (so the false-positive rate is $\mathbb{P}(T \mid D^c) = 0.10$). A randomly chosen person tests *positive*. What is the probability they actually have the disease?

■ **SOLUTION** Let D = "has disease" and T = "tests positive."
Step 1. Compute $\mathbb{P}(T)$ via total probability:

$$\begin{aligned}\mathbb{P}(T) &= \mathbb{P}(T \mid D)\mathbb{P}(D) + \mathbb{P}(T \mid D^c)\mathbb{P}(D^c) \\ &= (0.95)(0.01) + (0.10)(0.99) \\ &= 0.0095 + 0.0990 = 0.1085.\end{aligned}$$

Step 2. Apply Bayes' theorem:

$$\mathbb{P}(D \mid T) = \frac{(0.95)(0.01)}{0.1085} = \frac{0.0095}{0.1085} \approx 0.088.$$

Even after a positive test, there is only an 8.8% probability of disease. This counterintuitive result — called the **base-rate fallacy** when ignored — arises because the disease is rare (1% prevalence) and the false-positive rate (10%) swamps the true positives at low prevalence. ■

EXAMPLE 3.7 • Diagnosing the Cause of a System Outage

An IT team classifies outages into three mutually exclusive causes with prior probabilities:

$$\mathbb{P}(A_1) = 0.50, \quad \mathbb{P}(A_2) = 0.30, \quad \mathbb{P}(A_3) = 0.20,$$

where A_1 = network issue, A_2 = database issue, A_3 = application bug. A "high memory usage" alert (event B) fires. Likelihoods are:

$$\mathbb{P}(B \mid A_1) = 0.10, \quad \mathbb{P}(B \mid A_2) = 0.40, \quad \mathbb{P}(B \mid A_3) = 0.70.$$

Update the probability of each cause given the alert.

■ **SOLUTION** **Step 1.** Compute $\mathbb{P}(B)$ via total probability:

$$\begin{aligned}\mathbb{P}(B) &= (0.10)(0.50) + (0.40)(0.30) + (0.70)(0.20) \\ &= 0.05 + 0.12 + 0.14 = 0.31.\end{aligned}$$

Step 2. Apply Bayes' theorem to each cause.

Cause	Prior $\mathbb{P}(A_i)$	Likelihood $\mathbb{P}(B \mid A_i)$	Product $\mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$	Posterior $\mathbb{P}(A_i \mid B)$
Network (A_1)	0.50	0.10	0.05	0.161
Database (A_2)	0.30	0.40	0.12	0.387
App bug (A_3)	0.20	0.70	0.14	0.452
Total	1.00		0.31	1.000

Interpretation. Before the alert, the network issue was most probable (50%). After the alert fires, the application bug becomes most likely (45.2%) because application bugs have the highest memory-alert rate. The network cause drops from 50% to 16.1% because it rarely triggers memory alerts. This shift from prior to posterior is the essence of Bayesian updating. ■

GUIDED PRACTICE 3.3

A factory has three production lines. Line 1 produces 50% of parts with a 2% defect rate; Line 2 produces 30% with 5%; Line 3 produces 20% with 10%.

(a) Use the law of total probability to find the overall defect probability.

(b) A randomly chosen part is defective. Use Bayes' theorem to find the probability it came from Line 3.

(c) Which line becomes *most* probable given a defect?

Answers: (a) $(0.02)(0.50) + (0.05)(0.30) + (0.10)(0.20) = 0.010 + 0.015 + 0.020 = 0.045$. (b) $0.020/0.045 \approx 0.444$. (c) Line 3 (posterior $\approx$ 44.4%), despite producing only 20% of output.

EXERCISES (§3.2)

1. A company installs security cameras at its building. The probability that a camera's sensor is defective is 0.03. When a camera runs all day, a camera with a defective sensor has a 0.8 probability of losing its picture, while a camera with a normal sensor has only a 0.04 probability of losing its picture. Answer the following questions.

   (a) Find the probability that a randomly chosen camera loses its picture during operation. (Define all events clearly before solving.)

   (b) Find the conditional probability that a randomly chosen camera has a defective sensor, given that it has lost its picture.

2. A report states that the probability of the enemy launching an attack this month is 0.01. A radar detection system is used to identify incoming signals. When an attack occurs, the system detects an alarm with probability 0.98. When no attack occurs, the system still detects an alarm with probability 0.05. Answer the following questions.

(a) Find the probability that the radar detection system detects an alarm. (Define the relevant events before solving.)

(b) Given that the radar detection system has detected an alarm, find the probability that the enemy actually launched an attack.

3. Among first-year cadets, 90% are male and 10% are female. Among male cadets, 20% scored full marks on the midterm, while 21% of female cadets scored full marks. Answer the following.

(a) Find the probability that a randomly selected cadet scored full marks.

(b) Given that a randomly selected cadet scored full marks, what is the probability that the cadet is male?

4. According to an analysis by a national intelligence agency, a recent cyberattack was carried out by either Hacker Group A or Hacker Group B. The probabilities that the attack was conducted by Group A and Group B are 40% and 60%, respectively. After analyzing the attack logs, investigators found signs of unauthorized access in some cases. The probability of such signs appearing is 80% if Group A was responsible, and 30% if it was Group B. Answer the following.

(a) What is the probability that signs of unauthorized access appear in the attack logs? (Define the relevant events before solving the problem.)

(b) Given that signs of unauthorized access were found, what is the probability that the actual attacker was Group A?

5. In a certain battalion, 0.3% of firearms are defective. If soldiers used defective firearms, 94% experienced a malfunction during shooting. If soldiers used normal (non-defective) firearms, 2% experienced a malfunction during shooting. Answer the following questions.

(a) Calculate the probability that a soldier in this battalion experiences a malfunction during shooting. (Define the relevant events before solving the problem.)

(b) If a randomly selected soldier experiences a malfunction during shooting, find the probability that the soldier's firearm is defective.

6. A company uses battery packs for radios, but sometimes the batteries are defective. The probability that a battery pack is defective is 0.01. If the battery pack is defective, the probability that the radio does not work properly is 0.75. If the battery pack is not defective, the probability that the radio does not work properly is 0.02. Answer the following questions.

(a) Find the probability that a randomly chosen radio does not work properly. (Define the relevant events before solving the problem.)

(b) Find the conditional probability that the battery pack is defective, given that the radio does not work properly.

SUMMARY OF KEY CONCEPTS

- A **sample space** Ω lists all outcomes; events are subsets.
- The **Kolmogorov axioms** define a valid probability measure; all formulas follow from them.
- The **complement strategy** $\mathbb{P}(\text{at least one}) = 1 - \mathbb{P}(\text{none})$ simplifies many calculations.
- **Independence**: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$; knowledge of one event gives no information about the other.
- **Conditional probability** updates beliefs given partial information.

- **Bayes' theorem** reverses the conditioning: prior $\times$ likelihood $\rightarrow$ posterior.

CHAPTER 4

Random Variables

A **random variable** turns the outcome of a random experiment into a number, so that instead of listing events we can work with a single distribution. This chapter defines discrete and continuous random variables together with their probability mass and density functions, then summarizes a distribution through its **expected value** and **variance**. Finally we study two random variables together — their joint distribution, marginals, independence, covariance, and correlation — and show how means and variances behave under linear combinations.

4.1 Defining random variables

So far we have assigned probabilities to events in a sample space. In practice, however, we often summarize the outcome of a random experiment with a single number — the number of defectives in a batch, the waiting time until the next bus, the total repair cost of a failed component. A **random variable** formalises this idea.

DEFINITION 4.1 • Random Variable

A **random variable** (RV) X is a function that assigns a real number to each outcome in the sample space Ω.

- A **discrete** random variable takes values in a finite or countably infinite set $\{x_1, x_2, \ldots\}$.
- A **continuous** random variable takes values in an interval (or union of intervals) of real numbers.

We write $\mathbb{P}(X = x)$ for the probability that X takes the value x, and $\mathbb{P}(X \leq x)$ for the probability that X does not exceed x.

REMARK 4.1 • Notation convention

Capital letters (X, Y, Z) denote random variables; the corresponding lowercase letters (x, y, z) denote the specific values they may take. For example, $\mathbb{P}(X = x)$ is read "the probability that X equals x."

EXAMPLE 4.1 • Classifying Random Variables

- **Discrete:** Number of emails received in an hour; outcome of rolling a die ($\{1, 2, 3, 4, 5, 6\}$); number of defective units in a shipment of 100.
- **Continuous:** Time (in minutes) until a server responds to a request; weight (in kg) of a randomly selected soldier; battery life (in hours) of a device.

The distinction matters because it determines which mathematical tool — a sum or an integral — we use to compute probabilities.

4.1.1 Discrete random variables and the PMF

DEFINITION 4.2 • Probability Mass Function

The **probability mass function** (PMF) of a discrete random variable X is

$$f(x) = \mathbb{P}(X = x).$$

A valid PMF satisfies:

(i) $0 \leq f(x) \leq 1$ for all x.

(ii) $\sum_{\forall x} f(x) = 1$.

The probability that X falls in a set A is $\mathbb{P}(X \in A) = \sum_{x \in A} f(x)$.

EXAMPLE 4.2 • Counting Heads in Two Coin Tosses

Toss a fair coin twice and let X be the number of heads.

Sample space: $\Omega = \{HH, HT, TH, TT\}$, each equally likely.

The possible values of X are 0, 1, and 2:

- $X = 0$: outcome TT — one way out of four.
- $X = 1$: outcomes HT or TH — two ways out of four.
- $X = 2$: outcome HH — one way out of four.

PMF table:

x	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

Sanity check: $\sum_{\forall x} f(x) = f(0) + f(1) + f(2) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$.

From this PMF we can answer questions such as:

$$\mathbb{P}(X \geq 1) = f(1) + f(2) = \frac{2}{4} + \frac{1}{4} = \frac{3}{4},$$

or equivalently, using the complement rule, $\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \frac{1}{4} = \frac{3}{4}$.

4.1.2 Continuous random variables and the PDF

For a continuous random variable, every individual value has probability zero ($\mathbb{P}(X = x_0) = 0$ for any fixed x_0), so we describe probabilities over *intervals* instead.

DEFINITION 4.3 • Probability Density Function

The **probability density function** (PDF) of a continuous random variable X is a function $f(x) \geq 0$ such that for any $a < b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

A valid PDF satisfies:

(i) $f(x) \geq 0$ for all x.

(ii) $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1.$

Because individual points have measure zero, $\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b)$ for continuous RVs.

REMARK 4.2 • PDF is not a probability

Unlike the PMF, the PDF value $f(x_0)$ is *not* a probability — it can exceed 1. Probability is obtained only by integrating f over an interval. Think of $f(x)$ as describing the *density* of probability near x, analogous to mass per unit length.

EXAMPLE 4.3 • Uniform Waiting Time

A bus arrives at a stop at a uniformly random time within the next 60 minutes. Model the waiting time X (in minutes) with PDF

$$f(x) = \frac{1}{60}, \quad 0 \le x \le 60,$$

and $f(x) = 0$ otherwise.

■ **SOLUTION** **Validity:** $f(x) \ge 0$ and $\int_0^{60} \frac{1}{60}\,dx = 1$.
Probability of waiting more than 45 minutes:

$$\mathbb{P}(X > 45) = \int_{45}^{60} \frac{1}{60}\,dx = \frac{60 - 45}{60} = \frac{15}{60} = 0.25.$$

Probability of waiting between 10 and 30 minutes:

$$\mathbb{P}(10 \le X \le 30) = \int_{10}^{30} \frac{1}{60}\,dx = \frac{30 - 10}{60} = \frac{20}{60} \approx 0.333.$$

■

EXAMPLE 4.4 • Finding the Normalizing Constant

A continuous RV X has PDF

$$f(x) = c\,x(1 - x), \quad 0 \le x \le 1,$$

and $f(x) = 0$ otherwise, where c is an unknown constant.

■ **SOLUTION** **Step 1.** Impose $\int_{-\infty}^{\infty} f(x)\,dx = 1$:

$$\int_0^1 c\,x(1 - x)\,dx = c\int_0^1 (x - x^2)\,dx = c\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = c \cdot \frac{1}{6} = 1.$$

Step 2. Solve: $c = 6$.

Bonus — $\mathbb{P}(X \le \frac{1}{2})$:

$$\mathbb{P}\left(X \le \tfrac{1}{2}\right) = \int_0^{1/2} 6x(1 - x)\,dx = 6\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^{1/2} = 6\left(\frac{1}{8} - \frac{1}{24}\right) = 6 \cdot \frac{1}{12} = \frac{1}{2}.$$

By symmetry of f around $x = \frac{1}{2}$, this result is expected. ■

GUIDED PRACTICE 4.1

A continuous RV X has PDF $f(x) = ke^{-3x}$ for $x \geq 0$ and $f(x) = 0$ otherwise.

(a) Find the constant k.

(b) Find $\mathbb{P}(X > 1)$.

(c) Find $\mathbb{P}(0.5 < X < 2)$.

Answers: (a) $k = 3$. (b) $e^{-3} \approx 0.050$. (c) $e^{-1.5} - e^{-6} \approx 0.220$.

4.1.3 The cumulative distribution function

The PMF and PDF describe a distribution locally — value by value, or through its density. It is often more convenient to work with the *accumulated* probability up to a point, which is defined in the same way for discrete and continuous random variables alike.

DEFINITION 4.4 • Cumulative Distribution Function

The **cumulative distribution function** (CDF) of a random variable X is

$$F(x) = \mathbb{P}(X \leq x), \qquad -\infty < x < \infty.$$

It is obtained from the PMF or PDF by accumulation:

$$\underbrace{F(x) = \sum_{t \leq x} f(t),}_{\text{discrete}} \qquad \underbrace{F(x) = \int_{-\infty}^{x} f(t)\,dt\,.}_{\text{continuous}}$$

Every CDF satisfies:

(i) F is non-decreasing, with $0 \leq F(x) \leq 1$.

(ii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

(iii) F is right-continuous.

Interval probabilities follow directly:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

REMARK 4.3 • Recovering the PMF or PDF from the CDF

For a discrete random variable the CDF is a step function whose jump at each value x equals the probability there, $f(x) = F(x) - F(x^-)$. For a continuous random variable F is continuous and differentiable wherever f is continuous, with $f(x) = F'(x)$.

EXAMPLE 4.5 • CDF of the Two-Coin Toss

Return to X = number of heads in two tosses of a fair coin, with PMF $f(0) = \frac{1}{4}$, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$. Accumulating these probabilities gives the step function

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{4}, & 0 \le x < 1, \\ \frac{3}{4}, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

■ **SOLUTION** The jumps occur exactly at the possible values $0, 1, 2$, and each jump height equals the corresponding PMF value ($\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$). For instance,

$$\mathbb{P}(0 < X \le 2) = F(2) - F(0) = 1 - \tfrac{1}{4} = \tfrac{3}{4},$$

which matches $f(1) + f(2)$ found earlier. ■

EXAMPLE 4.6 • CDF of a Uniform Waiting Time

For the bus waiting time with PDF $f(x) = \frac{1}{60}$ on $[0, 60]$, integrating the density gives

$$F(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x}{60}, & 0 \le x \le 60, \\ 1, & x > 60. \end{cases}$$

■ **SOLUTION** Here F rises continuously from 0 to 1, and differentiating the middle piece recovers the density, $F'(x) = \frac{1}{60}$. The probability of waiting more than 45 minutes is

$$\mathbb{P}(X > 45) = 1 - F(45) = 1 - \tfrac{45}{60} = 0.25,$$

agreeing with the direct integral computed earlier. ■

EXERCISES (§4.1)

1. A discrete random variable X has PMF $f(x) = cx^2$ for $x = 1, 2, 3$ and $f(x) = 0$ otherwise.
   (a) Find the constant c.

(b) Find $E(X)$ and $\text{Var}(X)$.

(c) Find $\mathbb{P}(X > 1)$.

2. Let X be a continuous random variable with PDF $f(x) = 2e^{-2x}$ for $x > 0$ (and 0 otherwise).

(a) Verify that f is a valid PDF.

(b) Find $\mathbb{P}(X > 1)$ and $\mathbb{P}(0.5 < X < 2)$.

(c) Find the CDF $F(x)$.

3. A discrete random variable X has the PMF below.

x	0	1	2	3
$f(x)$	0.1	0.3	0.4	0.2

(a) Verify that this is a valid PMF.

(b) Find $\mathbb{P}(X \geq 2)$.

(c) Write the CDF $F(x)$ and use it to find $\mathbb{P}(1 < X \leq 3)$.

4.2 Expectation and variance

Knowing the full distribution of a random variable is ideal, but we often want concise numerical summaries. The two most important are the **expected value** (a measure of center) and the **variance** (a measure of spread).

4.2.1 Expected value

DEFINITION 4.5 • Expected Value

The **expected value** (or **mean**) of a random variable X is

$$\mu = E(X) = \begin{cases} \displaystyle\sum_{\forall x} x\, f(x) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} x\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

More generally, for any function g,

$$E[g(X)] = \begin{cases} \displaystyle\sum_{\forall x} g(x)\, f(x) & \text{(discrete),} \\ \displaystyle\int_{-\infty}^{\infty} g(x)\, f(x)\, dx & \text{(continuous).} \end{cases}$$

REMARK 4.4 • Interpretation

$E(X)$ is the long-run average value of X over many independent repetitions of the experiment — the *center of mass* of the distribution. It need not be a value that X can actually take (e.g. a

fair die has $E(X) = 3.5$).

EXAMPLE 4.7 • Expected Number of Heads

Using the PMF from Example 4.2:

SOLUTION

$$E(X) = 0 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{2}{4} + 2 \cdot \tfrac{1}{4} = 0 + \tfrac{1}{2} + \tfrac{1}{2} = 1.$$

On average, two fair coin tosses produce exactly one head.

Expected value of $Y = (X-1)^2$:

$$\begin{aligned} E[(X-1)^2] &= (0-1)^2 \cdot \tfrac{1}{4} + (1-1)^2 \cdot \tfrac{2}{4} + (2-1)^2 \cdot \tfrac{1}{4} \\ &= \tfrac{1}{4} + 0 + \tfrac{1}{4} = \tfrac{1}{2}. \end{aligned}$$

■

EXAMPLE 4.8 • Expected Waiting Time

For the uniform PDF $f(x) = 1/60$ on $[0, 60]$ (Example 4.3):

SOLUTION

$$E(X) = \int_0^{60} x \cdot \frac{1}{60}\, dx = \frac{1}{60} \cdot \frac{x^2}{2}\bigg|_0^{60} = \frac{1}{60} \cdot \frac{3600}{2} = 30 \text{ min}.$$

On average, we wait 30 minutes — exactly halfway through the interval, as symmetry demands. ■

LINEARITY OF EXPECTATION

For any random variable X and constants a, b:

$$E(aX + b) = a\,E(X) + b.$$

PROOF For the continuous case,

$$E(aX+b) = \int_{-\infty}^{\infty} (ax+b) f(x)\,dx = a\int_{-\infty}^{\infty} x f(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx = aE(X) + b.$$

The discrete case follows by replacing integrals with sums. ■

4.2.2 Variance and standard deviation

DEFINITION 4.6 • Variance and Standard Deviation

The **variance** of X measures the average squared deviation of X from its mean:

$$\sigma^2 = \mathrm{Var}(X) = E\Big[(X-\mu)^2\Big] = \begin{cases} \displaystyle\sum_{\forall x} (x-\mu)^2 f(x) & \text{(discrete)}, \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{(continuous)}. \end{cases}$$

The **standard deviation** is $\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$, which is measured in the same units as X.

COMPUTING FORMULA FOR VARIANCE

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2.$$

PROOF Expanding $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$ and applying linearity of expectation:

$$\mathrm{Var}(X) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2.$$

■

EXAMPLE 4.9 • Variance of the Coin-Toss RV

Continuing Example 4.7 with $\mu = 1$:

■ SOLUTION **Method 1 — direct definition:**

$$\mathrm{Var}(X) = (0-1)^2\cdot\tfrac{1}{4} + (1-1)^2\cdot\tfrac{2}{4} + (2-1)^2\cdot\tfrac{1}{4} = \tfrac{1}{4} + 0 + \tfrac{1}{4} = \tfrac{1}{2}.$$

Method 2 — computing formula:

$$E(X^2) = 0^2\cdot\tfrac{1}{4} + 1^2\cdot\tfrac{2}{4} + 2^2\cdot\tfrac{1}{4} = 0 + \tfrac{1}{2} + 1 = \tfrac{3}{2}.$$

$$\mathrm{Var}(X) = \tfrac{3}{2} - 1^2 = \tfrac{1}{2}.$$

The standard deviation is $\sigma = \sqrt{1/2} \approx 0.707$. ■

EXAMPLE 4.10 • Mean and Variance of a Continuous RV

Let X be a waiting time with $f(x) = 1$ on $[0, 1]$ (a standard uniform).

■ SOLUTION **Mean:**

$$E(X) = \int_0^1 x\cdot 1\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}.$$

Second moment:

$$E(X^2) = \int_0^1 x^2\cdot 1\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}.$$

Variance:

$$\mathrm{Var}(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

The standard deviation is $\sigma = 1/\sqrt{12} \approx 0.289$. ■

PROPERTIES OF VARIANCE

For any random variable X and constants a, b:

(i) $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.

(ii) $\mathrm{Var}(X) \geq 0$, with equality if and only if X is constant.

■ PROOF (i) Let $\mu' = E(aX + b) = a\mu + b$. Then $aX + b - \mu' = a(X - \mu)$, so

$$\mathrm{Var}(aX + b) = E\left[a^2(X-\mu)^2\right] = a^2\,\mathrm{Var}(X).$$

(ii) $\mathrm{Var}(X) = E[(X-\mu)^2] \geq 0$ since $(X-\mu)^2 \geq 0$. ■

GUIDED PRACTICE 4.2

A quality-control inspector scores parts from 0 to 10. The score X has PMF $f(0) = 0.1$, $f(5) = 0.6$, $f(10) = 0.3$.

(a) Find $E(X)$.

(b) Find $\mathrm{Var}(X)$ using the computing formula.

(c) If the score is linearly transformed as $Y = 2X - 3$, find $E(Y)$ and $\mathrm{SD}(Y)$.

Answers: (a) $0(0.1) + 5(0.6) + 10(0.3) = 6$. (b) $E(X^2) = 0(0.1) + 25(0.6) + 100(0.3) = 45$; $\mathrm{Var}(X) = 45 - 36 = 9$. (c) $E(Y) = 2(6) - 3 = 9$; $\mathrm{SD}(Y) = 2\sqrt{9} = 6$.

EXERCISES (§4.2)

1. Suppose $E(X) = 4$ and $\mathrm{Var}(X) = 9$. Find:
   (a) $E(2X + 5)$;
   (b) $\mathrm{Var}(2X + 5)$;
   (c) $\mathrm{SD}(3 - X)$.
2. A discrete random variable X takes the values $-1, 0, 2$ with probabilities $0.5, 0.3, 0.2$. Compute $E(X)$ and $E(X^2)$, then use the computing formula $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$ to find $\mathrm{Var}(X)$.
3. A game costs \$2 to play: you roll a fair die and win a number of dollars equal to the face shown.
   (a) Find the expected net gain per play.
   (b) Is the game favorable to the player? Justify using the expected value.

4.3 Joint distributions

Many practical problems involve *two or more* random variables simultaneously. For instance, an engineer might record both the load X on a bridge and the resulting deflection Y; a logistics manager might track both the number of packages arriving X and the number of drivers available Y. A **joint distribution** describes how two random variables behave together.

4.3.1 Joint PMF and PDF

DEFINITION 4.7 • Joint Distribution

For two discrete RVs X and Y, the **joint PMF** is

$$f(x, y) = \mathbb{P}(X = x,\ Y = y).$$

It satisfies $f(x, y) \geq 0$ and $\sum_{\forall x} \sum_{\forall y} f(x, y) = 1$.

For two continuous RVs, the **joint PDF** $f(x,y) \geq 0$ satisfies $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$, and

$$\mathbb{P}(a < X < b,\ c < Y < d) = \int_a^b \int_c^d f(x,y)\,dy\,dx.$$

EXAMPLE 4.11 • Cars and Television Sets

A survey of suburban households recorded the number of cars X owned and the number of television sets Y owned. The joint PMF is tabulated below.

$X \backslash Y$	1	2	3	4
1	0.10	0.00	0.10	0.00
2	0.30	0.00	0.10	0.20
3	0.00	0.20	0.00	0.00

All entries are non-negative and sum to 1. For instance, $\mathbb{P}(X = 2, Y = 4) = 0.20$: 20% of households own exactly 2 cars and 4 televisions.

4.3.2 Marginal distributions

Given a joint distribution, we recover each variable's individual distribution by *summing out* (discrete) or *integrating out* (continuous) the other variable.

DEFINITION 4.8 • Marginal Distribution

For discrete RVs with joint PMF $f(x,y)$:

$$f_X(x) = \sum_{\forall y} f(x,y), \qquad f_Y(y) = \sum_{\forall x} f(x,y).$$

For continuous RVs with joint PDF $f(x,y)$:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx.$$

EXAMPLE 4.12 • Marginals for Cars and TVs

Continuing Example 4.11, we sum each row for f_X and each column for f_Y.

$X\backslash Y$	1	2	3	4	$f_X(x)$
1	0.10	0.00	0.10	0.00	0.20
2	0.30	0.00	0.10	0.20	0.60
3	0.00	0.20	0.00	0.00	0.20
$f_Y(y)$	0.40	0.20	0.20	0.20	1.00

For instance, $f_X(2) = 0.30 + 0.00 + 0.10 + 0.20 = 0.60$: 60% of households own exactly 2 cars regardless of how many TVs they have. We read the marginals directly from the row and column totals.

4.3.3 Independence of two random variables

DEFINITION 4.9 • Independence of RVs

Random variables X and Y are **independent** if their joint distribution factors into the product of their marginals:

$$f(x, y) = f_X(x)\, f_Y(y) \quad \text{for all } x, y.$$

This directly extends the event-independence condition $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ to random variables.

EXAMPLE 4.13 • Checking Independence via Factorisation

Consider the joint PDF

$$f(x, y) = \frac{3x - y}{9}, \quad 1 < x < 3,\ 1 < y < 2,$$

and $f(x, y) = 0$ otherwise. Are X and Y independent?

■ SOLUTION **Step 1.** Find the marginal of X:

$$f_X(x) = \int_1^2 \frac{3x - y}{9}\, dy = \frac{1}{9}\left[3xy - \frac{y^2}{2}\right]_1^2 = \frac{1}{9}\left(3x - \frac{3}{2}\right) = \frac{6x - 3}{18}, \quad 1 < x < 3.$$

Step 2. Find the marginal of Y:

$$f_Y(y) = \int_1^3 \frac{3x - y}{9}\, dx = \frac{1}{9}\left[\frac{3x^2}{2} - xy\right]_1^3 = \frac{12 - 2y}{9}, \quad 1 < y < 2.$$

Step 3. Check factorisation:

$$f_X(x)\, f_Y(y) = \frac{6x-3}{18} \cdot \frac{12-2y}{9} = \frac{(6x-3)(12-2y)}{162}.$$

This does not simplify to $\dfrac{3x-y}{9}$ in general, so X and Y are **not independent**.

Intuitively, the numerator $3x - y$ mixes x and y in a way that cannot be separated into a product of a function of x alone and a function of y alone. ■

4.3.4 Covariance and correlation

When two random variables are dependent, **covariance** and **correlation** quantify the direction and strength of their linear association.

DEFINITION 4.10 • Covariance

The **covariance** of X and Y is

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)\,E(Y),$$

where $E(XY)$ is computed from the joint distribution:

$$E(XY) = \begin{cases} \displaystyle\sum_{\forall x}\sum_{\forall y} xy\, f(x,y) & \text{(discrete)}, \\ \displaystyle\int\int xy\, f(x,y)\, dx\, dy & \text{(continuous)}. \end{cases}$$

If $\mathrm{Cov}(X,Y) > 0$, larger values of X tend to accompany larger values of Y; if $\mathrm{Cov}(X,Y) < 0$, they tend to move in opposite directions.

! CAUTION

Zero covariance does not imply independence. $\mathrm{Cov}(X,Y) = 0$ rules out only a *linear* relationship. X and Y could still be strongly dependent in a non-linear way (e.g. $Y = X^2$ with X symmetric around zero has $\mathrm{Cov}(X,Y) = 0$ but Y is entirely determined by X). Independence, however, always implies $\mathrm{Cov}(X,Y) = 0$.

DEFINITION 4.11 • Correlation

The **correlation** between X and Y is the scale-free version of covariance:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

It always satisfies $-1 \le \rho \le 1$.

- $\rho > 0$: positive linear relationship.
- $\rho < 0$: negative linear relationship.
- $\rho = 0$: no linear relationship.
- $|\rho| = 1$: perfect linear relationship.

EXAMPLE 4.14 • Computing Covariance from a Joint PMF

Using the joint and marginal PMFs from Example 4.12:

■ **SOLUTION** **Step 1.** Compute $E(X)$ and $E(Y)$:

$$E(X) = 1(0.20) + 2(0.60) + 3(0.20) = 2.00.$$

$$E(Y) = 1(0.40) + 2(0.20) + 3(0.20) + 4(0.20) = 2.20.$$

Step 2. Compute $E(XY)$ by multiplying each cell's xy product by its probability and summing:

$$\begin{aligned} E(XY) &= (1)(1)(0.10) + (1)(3)(0.10) + (2)(1)(0.30) + (2)(3)(0.10) \\ &\quad + (2)(4)(0.20) + (3)(2)(0.20) \\ &= 0.10 + 0.30 + 0.60 + 0.60 + 1.60 + 1.20 = 4.40. \end{aligned}$$

Step 3. Apply the shortcut formula:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 4.40 - (2.00)(2.20) = 0.$$

The covariance is zero, yet X and Y are not independent (we can verify $f(1,1) = 0.10 \neq f_X(1)f_Y(1) = (0.20)(0.40) = 0.08$). This illustrates that $\text{Cov}(X, Y) = 0$ is a necessary but not sufficient condition for independence. ■

4.3.5 Linear combinations of random variables

In practice we often work with sums and weighted combinations of random variables — total cost, total time, portfolio return. The following results make such calculations straightforward.

EXPECTATION OF A LINEAR COMBINATION

For any two random variables X and Y and constants a, b:

$$E(aX + bY) = a\,E(X) + b\,E(Y).$$

More generally, for $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$:

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n).$$

No independence assumption is required.

VARIANCE OF A LINEAR COMBINATION

If X and Y are **independent**:

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y).$$

Without independence, the general formula is:

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

! CAUTION

Doubling one observation vs. summing two. If X_1 and X_2 are *independent* copies of the same distribution, then

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\mathrm{Var}(X).$$

But writing $2X$ means doubling the *same* observation, and a variable is perfectly correlated with itself, so

$$\mathrm{Var}(2X) = 4\mathrm{Var}(X).$$

Always clarify: is this one observation scaled up, or two separate independent observations added together?

EXAMPLE 4.15 • Procurement Cost

A unit buys 5 notebooks and 4 pens each month. The price per notebook X has mean \$10 and standard deviation \$1.50; the price per pen Y has mean \$15 and standard deviation \$2.00. Assume X and Y are independent.

■ SOLUTION **Expected total cost:**

$$E(5X + 4Y) = 5\,E(X) + 4\,E(Y) = 5(10) + 4(15) = 50 + 60 = \$110.$$

Variance of total cost:

$$\operatorname{Var}(5X + 4Y) = 25\operatorname{Var}(X) + 16\operatorname{Var}(Y) = 25(2.25) + 16(4.00) = 56.25 + 64.00 = 120.25.$$

Standard deviation of total cost:

$$\operatorname{SD}(5X + 4Y) = \sqrt{120.25} \approx \$10.97.$$

The unit expects to spend \$110 per month on these supplies, with a standard deviation of about \$10.97. ■

GUIDED PRACTICE 4.3

A network packet traverses three independent hops. Hop i introduces delay X_i (ms) with mean μ_i and variance σ_i^2: $\mu_1 = 5, \sigma_1^2 = 4$; $\mu_2 = 3, \sigma_2^2 = 1$; $\mu_3 = 7, \sigma_3^2 = 9$.

(a) Find the expected total delay.

(b) Find the standard deviation of the total delay.

(c) Convert the results to seconds (multiply by 0.001) and state E and SD.

Answers: (a) $5+3+7 = 15$ ms. (b) $\sqrt{4+1+9} = \sqrt{14} \approx 3.74$ ms. (c) $E = 0.015$ s, SD ≈ 0.00374 s.

EXERCISES (§4.3)

1. The joint probability mass function of two random variables X and Y and its table are given below. Answer the following questions.

$$f(x,y) = \begin{cases} \dfrac{x+y}{9}, & x = 0,1,2,\ y = 0,1 \\ 0, & \text{otherwise} \end{cases}$$

$x \backslash y$	0	1
0	0	1/9
1	1/9	2/9
2	2/9	3/9

(a) Find the marginal probability mass function $f_X(x)$ of X.

(b) The marginal probability mass function $f_Y(y)$ of Y is given below. Find the variance $\operatorname{Var}(Y)$. (You may use $E(Y) = \frac{2}{3}$.)

$$f_Y(y) = \begin{cases} 1/3, & y = 0 \\ 2/3, & y = 1 \\ 0, & \text{otherwise.} \end{cases}$$

(c) Find $E(2X - 3Y + 4)$.

(d) Determine whether X and Y are independent.

2. The joint probability density function $f(x, y)$ of two random variables X and Y and the marginal probability density function $f_Y(y)$ of Y are given by

$$f(x,y) = \begin{cases} e^{-x}, & x > 0,\ 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}, \quad f_Y(y) = \begin{cases} 1, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Answer the following questions.

(a) Find the marginal probability density function $f_X(x)$ of X. (State the range of possible values of X).

(b) Determine whether X and Y are independent.

(c) Find the expected value $E(Y)$.

(d) Find the covariance $\text{Cov}(X, Y)$. You may use $E(XY) = \frac{1}{2}$ and $E(X) = 1$.

3. The joint probability mass function (pmf) of two random variables X and Y is given by

$$f(x,y) = \begin{cases} \dfrac{x^2 y}{42}, & x = 1,2,3, \quad y = 1,2, \\ 0, & \text{otherwise.} \end{cases}$$

Answer the following.

(a) Find the marginal pmf of X, $f_X(x)$, and the marginal pmf of Y, $f_Y(y)$.

(b) Determine whether the two random variables X and Y are independent.

(c) Find the expectation of Y, $E(Y)$.

(d) Find the variance of Y, $\text{Var}(Y)$.

4. Alex and Ben each depart from home to a meeting point at the same time. The time taken (in minutes) for Alex is X and for Ben is Y, and their joint probability density function is

$$f(x,y) = \begin{cases} \dfrac{1}{200}, & 35 \le x \le 55,\ 30 \le y \le 40, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the probability that Alex arrives earlier than Ben, $\mathbb{P}(X < Y)$.

(b) Suppose Alex and Ben make this trip $n = 30$ times. Out of the 30 trips, find the expected number of times that Alex arrives earlier than Ben. (Assume that each trip is independent.)

5. The joint probability density function (pdf) of two random variables X and Y and the marginal pdf of X are given by

$$f(x,y) = \begin{cases} 4xy, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{otherwise,} \end{cases} \qquad f_X(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Answer the following questions.

(a) Find the marginal probability density function of Y, $f_Y(y)$.

(b) Find the expectation of X, $E(X)$.

(c) Determine whether the two random variables X and Y are independent.

SUMMARY OF KEY CONCEPTS

- A **random variable** maps outcomes to numbers; its distribution is described by a PMF (discrete) or PDF (continuous).
- $E(X)$ is the probability-weighted mean; $\text{Var}(X) = E(X^2) - [E(X)]^2$ is its spread.

- The **joint distribution** of (X, Y) encodes their relationship; marginals are obtained by summing or integrating out the other variable.
- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ measures linear association; $\rho(X, Y) = \text{Cov}/\sqrt{\text{Var}(X)\text{Var}(Y)}$ scales it to $[-1, 1]$.
- For linear combinations: $E(aX + bY) = aE(X) + bE(Y)$ always; $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$ when X, Y are independent.

CHAPTER 5

Distributions

Chapter 4 introduced random variables and the rules that govern them. We now meet the handful of **named distributions** that recur everywhere in statistics and machine learning — models flexible enough to describe real data yet simple enough to compute with. This chapter studies three: the **uniform**, the **normal**, and the **binomial**. The first two are *continuous* (probability is the *area* under a density curve); the third is *discrete* (probability is a *sum* of point masses). Each is specified by one or two **parameters**, and from those parameters we can read off its center, its spread, and the probability of any event of interest.

5.1 Uniform distribution

The simplest continuous model is one in which every value in an interval is equally likely. It is the natural description of "complete ignorance" within a known range, and it is the foundation of computer simulation: nearly every random quantity a program generates begins life as a draw from the **standard uniform** distribution on $[0, 1]$.

DEFINITION 5.1 • Continuous uniform distribution

A random variable X is **uniform** on the interval $[a, b]$, written $X \sim \text{Uniform}(a, b)$, if its probability density function is constant on that interval:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

Because the density is a rectangle of height $1/(b-a)$ and width $b-a$, its total area is 1, as every density must be. Probabilities are areas of sub-rectangles, which makes them trivial to compute: for any $a \le c \le d \le b$,

$$\mathbb{P}(c \le X \le d) = \frac{d-c}{b-a}.$$

Integrating $xf(x)$ and $(x-\mu)^2 f(x)$ over $[a, b]$ (Definitions 4.5 and 4.6) gives the mean and variance,

$$\mathbb{E}[X] = \frac{a+b}{2}, \qquad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

EXAMPLE 5.1 • Quantization error

To shrink a neural network, its 32-bit weights are often **quantized** to a coarse grid of integers. Rounding each weight to the nearest grid point introduces an error that, to good approximation, is uniform on $(-\frac{1}{2}, \frac{1}{2})$ in units of one grid step. Here $a = -\frac{1}{2}$ and $b = \frac{1}{2}$, so the density is $f(x) = 1$ on that interval (Figure 5.1).

The error has mean $\mathbb{E}[X] = 0$ — rounding is unbiased — and variance $\text{Var}(X) = 1/12 \approx 0.083$, so its standard deviation is $\sqrt{1/12} \approx 0.289$ steps. The probability that a given weight's rounding error lands between 0.1 and 0.4 is simply the shaded area,

$$\mathbb{P}(0.1 \le X \le 0.4) = \frac{0.4 - 0.1}{1} = 0.3.$$

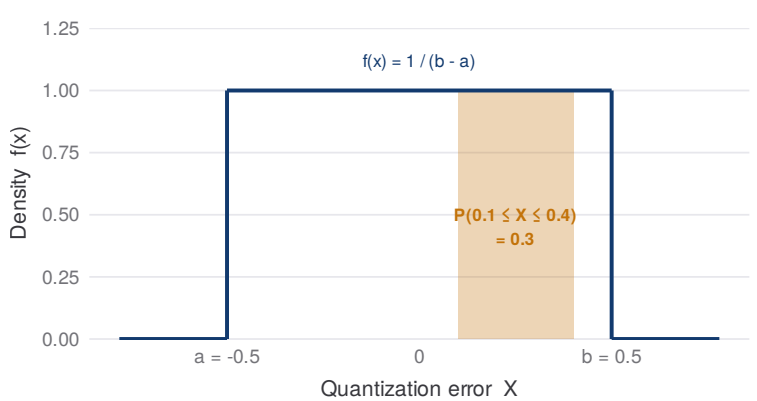


Figure 5.1. Density of the quantization error $X \sim \text{Uniform}(-0.5, 0.5)$. Probability is area: the shaded rectangle has area 0.3.

REMARK 5.1 • Uniformity powers simulation

A draw $U \sim \text{Uniform}(0,1)$ can be turned into a draw from *any* distribution by the inverse-CDF method, which is why the standard uniform sits beneath random shuffling, dropout masks, and weight initialisation. We rely on it implicitly every time we set a random seed.

GUIDED PRACTICE 5.1

A background job is scheduled to start at a uniformly random instant within a 60-second polling window, so its start time $S \sim \text{Uniform}(0, 60)$ seconds. What is the probability it starts within the first 15 seconds, and what is its mean start time?
Answer: $\mathbb{P}(S \le 15) = 15/60 = 0.25$, and $\mathbb{E}[S] = (0 + 60)/2 = 30$ seconds.

EXERCISES (§5.1)

1. Jisoo's waiting time for bus X, measured in minutes, is uniformly distributed on the interval $(0, 15)$.
   (a) Find the probability density function(pdf) of X.
   (b) What is the probability that Jisoo's waiting time X is less than 3 minutes?
   (c) Find the expectation and variance of X.
2. In a shooting exercise, a cadet aims at a target, and the distance X (cm) from the hit point to the center follows a uniform distribution on $(0, 10)$. The score Y is 10 points if $X < 4$, and 5 points if $4 \le X < 10$. Answer the following questions.
   (a) Find the probability density function $f_X(x)$ and the mean $E(X)$.
   (b) Find $\mathbb{P}(Y = 10)$ and $\mathbb{P}(Y = 5)$.
   (c) Find the mean $E(Y)$.
3. The lifetime (in hours) of an electronic device follows a uniform distribution, $X \sim U(0, 30)$. Answer the following.
   (a) Find the cumulative distribution function (CDF) $F_X(x) = \mathbb{P}(X \le x)$ for $0 < x < 30$.
   (b) Find the probability that at least two out of five (independently chosen) devices last more than 10 hours.

5.2 Normal distribution

The **normal** (or **Gaussian**) distribution is the most important model in all of statistics. Its symmetric, bell-shaped curve describes measurement errors, aggregates of many small effects, and — through the central limit theorem of a later chapter — the behavior of sample averages. In machine learning it is the default model for prediction errors and the basis of countless algorithms.

DEFINITION 5.2 • Normal distribution

A random variable X is **normal** with mean μ and standard deviation σ, written $X \sim N(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

The parameter μ locates the center of the curve and σ controls its width (Figure 5.2).

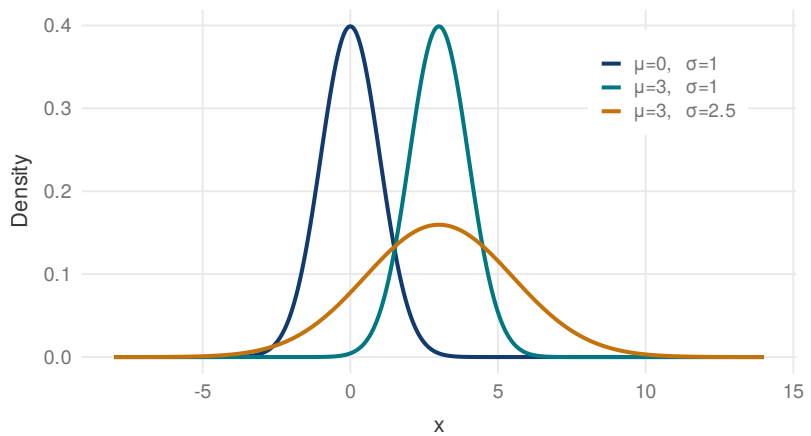


Figure 5.2. Normal densities for three parameter choices. Changing μ shifts the curve left or right; increasing σ makes it wider and flatter while preserving total area 1.

5.2.1 Standardizing with Z-scores

Every normal distribution is a rescaled copy of one special case, the **standard normal** $N(0,1)$. Subtracting the mean and dividing by the standard deviation converts any normal value into a standard one.

DEFINITION 5.3 • Z-score

The **Z-score** of an observation x from a distribution with mean μ and standard deviation σ is

$$z = \frac{x-\mu}{\sigma}.$$

It records how many standard deviations x lies above (if positive) or below (if negative) the mean. If $X \sim N(\mu, \sigma^2)$ then $Z = (X-\mu)/\sigma \sim N(0,1)$.

Z-scores make values from different distributions directly comparable (Figure 5.3).

EXAMPLE 5.2 • Comparing across scales

A pixel intensity of 170 is observed in two image datasets. In dataset A intensities follow

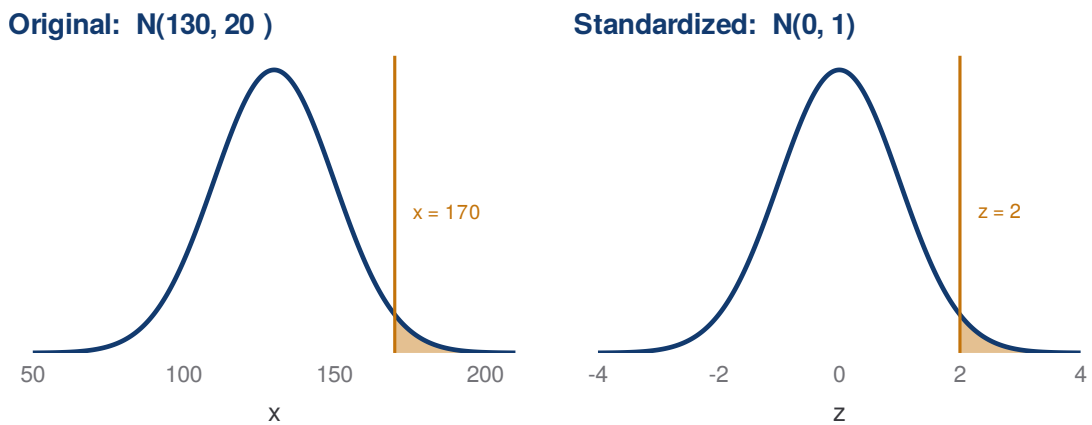


Figure 5.3. Standardizing maps a value x from $N(\mu, \sigma^2)$ (left) to its Z-score z under the standard normal $N(0, 1)$ (right). The two shaded tail areas are equal — standardizing changes the scale, not the probability.

$N(130, 20^2)$, and in dataset B they follow $N(100, 40^2)$. Which dataset finds 170 more unusual? Standardizing,

$$z_A = \frac{170 - 130}{20} = 2.0, \qquad z_B = \frac{170 - 100}{40} = 1.75.$$

Although 170 is farther from B's mean in raw units, it is more standard deviations from A's mean, so it is more unusual relative to dataset A.

5.2.2 Properties and the 68–95–99.7 rule

The normal density is symmetric about μ, so the mean, median, and mode coincide there, and the total area under the curve is 1. A remarkably useful consequence is a fixed relationship between standard deviations and probability.

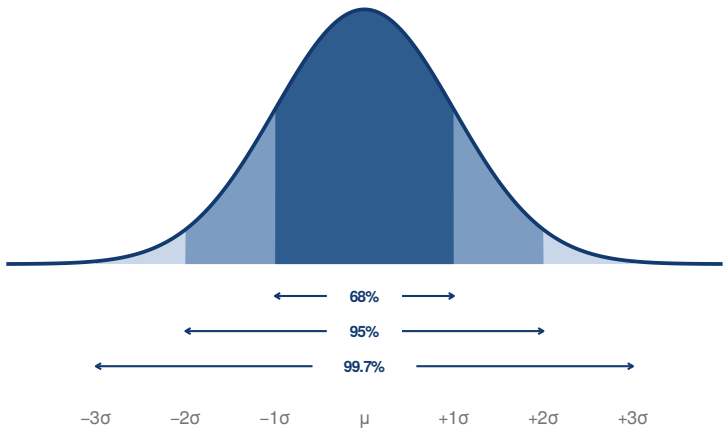


Figure 5.4. The empirical rule. The shaded bands within $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ contain about 68%, 95%, and 99.7% of the distribution.

THE 68–95–99.7 (EMPIRICAL) RULE

For any normal distribution, approximately

$$68\%, \quad 95\%, \quad \text{and} \quad 99.7\%$$

of the probability lies within 1, 2, and 3 standard deviations of the mean, respectively (Figure 5.4).

5.2.3 Finding areas and percentiles

For probabilities other than the round numbers of the empirical rule we use the **standard normal cumulative distribution function**

$$\Phi(z) = \mathbb{P}(Z \le z),$$

which has no closed form but is built into every statistical tool. In **R** the value $\Phi(z)$ is `pnorm(z)`, and the reverse question — which cutoff has a given area below it — is answered by `qnorm`. Historically the same values were read from a printed Z-table.

EXAMPLE 5.3 • A latency service-level objective

A cached endpoint responds in a time $T \sim N(50, 8^2)$ milliseconds, and the service-level objective (SLO) is to stay under 70 ms. What fraction of requests miss it? Standardize the cutoff and use the CDF:

$$\mathbb{P}(T > 70) = 1 - \Phi\left(\frac{70 - 50}{8}\right) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.$$

About 0.6% of requests exceed the budget (Figure 5.5, left).

A closely related task is to find a **percentile**: the value below which a given fraction of the distribution falls. The 95th percentile is the 0.95 **quantile**, found by inverting the CDF. Since $\Phi^{-1}(0.95) = 1.645$ (that is, `qnorm(0.95)` = 1.645),

$$p_{95} = \mu + 1.645\,\sigma = 50 + 1.645(8) = 63.2 \text{ ms}.$$

Ninety-five percent of requests finish within 63.2 ms (Figure 5.5, right). Engineers report exactly this "p95 latency" to summarize tail performance.

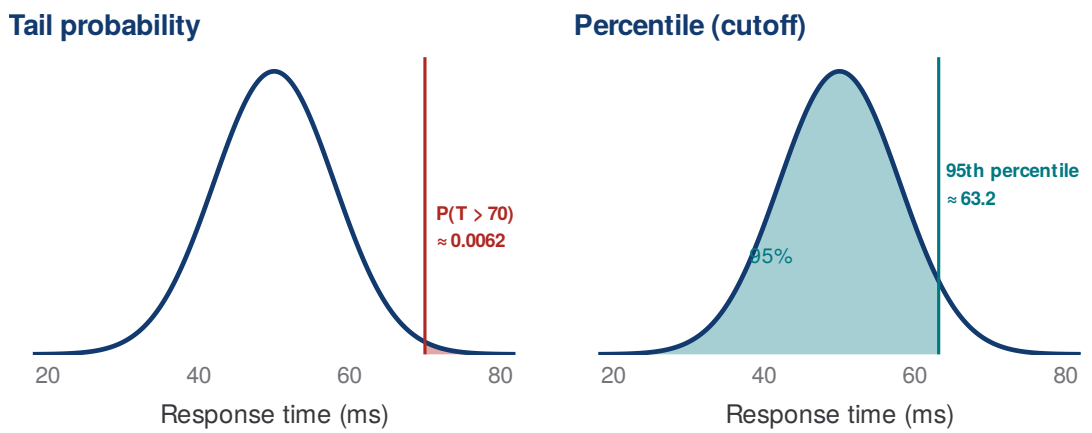


Figure 5.5. Left: the tail probability $\mathbb{P}(T > 70)$ for $T \sim N(50, 8^2)$. Right: the 95th percentile, the cutoff with 95% of the area to its left.

! CAUTION

Real data are at best *approximately* normal. Strongly skewed or heavy-tailed quantities — such as the raw inference latency of Chapter 2 — can be badly misdescribed by a normal model. Always check the shape (for example with a histogram) before trusting normal-based probabilities.

GUIDED PRACTICE 5.2

Scores produced by a model are approximately $N(70, 10^2)$. Using the empirical rule only, what fraction of scores fall between 60 and 90?
Answer: 60 is 1σ below and 90 is 2σ above the mean. That range holds the 34% between μ and -1σ plus the 47.5% between μ and $+2\sigma$, totalling about 81.5%.

EXERCISES (§5.2)

1. The battery life of portable radios produced by a company follows a normal distribution with mean 45 hours and standard deviation 3 hours. When 16 units are randomly selected, find the probability that the sample mean battery life is in between 44 hours and 47 hours. (Be sure to define your random variable before solving.)

2. At a factory, the weight (in grams) of fertilizer produced by Machine A follows a normal distribution with mean 100 g and standard deviation 3 g, while the weight of fertilizer produced by Machine B follows a normal distribution with mean 95 g and standard deviation 4 g. One product is randomly selected from each machine. Let X denote the weight (in grams) of fertilizer from Machine A and Y denote the weight (in grams) from Machine B. Find $\mathbb{P}(X \geq Y + 17.5)$.
3. The average daily high temperature in June in LA is 77°F with a standard deviation of 5°F. Suppose that the temperatures in June closely follow a normal distribution.
   (a) What is the probability of observing an 83°F temperature or higher in LA during a randomly chosen day in June?
   (b) How cool are the coldest 10% of the days (days with lowest high temperature) during June in LA?
4. At a smart farm, the seed size of a crop follows a normal distribution with mean μ and variance σ^2. Let X be the seed size (mm). It is known that $\mathbb{P}(X \geq 5) = 0.5$ and $\mathbb{P}(X \geq 8.92) = 0.025$. Answer the following.
   (a) Find $E(X)$ and $\text{Var}(X)$.
   (b) What is the cutoff for the lowest 2.5% of seed sizes?
5. Final grades in "Introduction to Statistics" are determined by combining 40% of the midterm and 60% of the final exam scores. The midterm scores follow $N(60, 10^2)$ and the final exam scores follow $N(70, 5^2)$, and the two scores are independent. Answer the following.
   (a) Describe the probability distribution of the final grade.
   (b) A cadet scored 70 on the midterm and 80 on the final exam. What percentile is their final grade?
   (c) If the top 20% of students receive an A grade, what is the minimum score required for an A?
6. At a university, final grades in a statistics course are based on a weighted average: 30% quiz (Q), 30% midterm (M), and 40% final exam (F). The quiz score Q follows $N(80, 5^2)$, the midterm score M follows $N(60, 20^2)$, and the final exam score F follows $N(70, 5^2)$. Assume the three scores are independent. The final weighted score is $W = 0.3\,Q + 0.3\,M + 0.4\,F$.
   (a) Find the distribution of the final score W.
   (b) A student scored 90 on the quiz, 80 on the midterm, and 85 on the final exam. What proportion of students have a final score W higher than this student? (Use $\mathbb{P}(Z > 2.3077) \approx 0.0105$.)
7. Suppose the time required to complete a task follows a normal distribution with mean 30 minutes and standard deviation 3 minutes.
   (a) Find the probability that the task completion time falls between 27 minutes and 36 minutes. (Use $\mathbb{P}(Z \leq -2) = 0.0228$ and $\mathbb{P}(Z \leq -1) = 0.1587$.)
   (b) What is the cutoff for the top 20% of task completion times? (Use $z_{0.20} = 0.8416$.)

5.3 Binomial distribution

Many questions reduce to *counting successes* in a fixed number of independent attempts: how many of n images a classifier labels correctly, how many of n requests succeed, how many of n users click. When the attempts are independent and each succeeds with the same probability, the count follows a **binomial** distribution. A single such attempt — one trial with only two possible outcomes — is described by the simplest discrete model of all, the **Bernoulli** distribution, from which the binomial is built by repetition.

5.3.1 The Bernoulli distribution

DEFINITION 5.4 • Bernoulli distribution

A random variable X has the **Bernoulli distribution** with **success probability** $p \in [0,1]$, written $X \sim \text{Bernoulli}(p)$, if it takes only the value 1 ("success") or 0 ("failure"), with

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$

Equivalently, its probability mass function is

$$\mathbb{P}(X = x) = p^x (1-p)^{1-x}, \qquad x \in \{0, 1\}.$$

Such an X is called an **indicator** variable, because it records whether a single event occurs. Its mean and variance follow directly from the definition of expectation: since X equals 1 with probability p and 0 otherwise — and $X^2 = X$ for a 0/1 variable —

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p, \qquad \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p).$$

The variance $p(1-p)$ is largest at $p = \frac{1}{2}$ — a fair coin is the least predictable trial — and shrinks to 0 as p approaches 0 or 1, where the outcome is nearly certain.

EXAMPLE 5.4 • A single classification

The classifier of Chapter 2 labels one image correctly with probability $p = 0.865$. Let $X = 1$ if the label is correct and $X = 0$ otherwise. Then $X \sim \text{Bernoulli}(0.865)$, with $\mathbb{E}[X] = 0.865$ and $\text{Var}(X) = 0.865 \times 0.135 \approx 0.117$. Evaluating n such images independently and counting the correct labels is exactly the binomial experiment of the next subsection.

REMARK 5.2 • From one trial to many

A binomial count is a sum of independent Bernoulli trials: if $X_1, \ldots, X_n$ are independent Bernoulli(p) variables, then $X = X_1 + \cdots + X_n \sim B(n, p)$. Adding the n identical means p and variances $p(1-p)$ is precisely why the binomial mean and variance work out to np and $np(1-p)$ below.

5.3.2 The binomial distribution

DEFINITION 5.5 • Binomial distribution

Suppose an experiment consists of n independent trials, each a "success" with probability p and a "failure" with probability $1 - p$. The number of successes X has the **binomial**

distribution $X \sim B(n, p)$, with probability mass function

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

where the **binomial coefficient** $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$ counts the number of distinct ways to arrange k successes among the n trials.

The factor $p^k(1-p)^{n-k}$ is the probability of one specific sequence with k successes; multiplying by $\binom{n}{k}$ adds up all such sequences.

EXAMPLE 5.5 • Correct predictions in a test batch

Recall the classifier of Chapter 2, which labels each image correctly with probability $p = 0.865$. If we evaluate $n = 20$ fresh images and assume independent outcomes, the number correct is $X \sim B(20, 0.865)$ (Figure 5.6). The probability of a perfect score is

$$\mathbb{P}(X = 20) = \binom{20}{20} (0.865)^{20} (0.135)^{0} = 0.865^{20} \approx 0.055,$$

and summing the top three masses gives $\mathbb{P}(X \geq 18) \approx 0.481$ — the model gets at least 18 of 20 right about half the time.

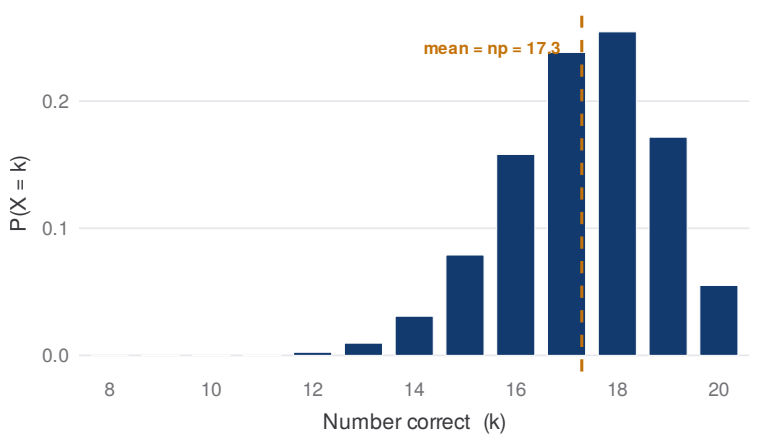


Figure 5.6. Binomial pmf for $X \sim B(20, 0.865)$, the number of correct predictions in 20 images. The dashed line marks the mean $np = 17.3$.

5.3.3 Expected value and variability

The mean and variance of a binomial follow from adding up the n identical trials.

MEAN AND STANDARD DEVIATION OF A BINOMIAL

If $X \sim B(n, p)$, then

$$\mathbb{E}[X] = np, \qquad \mathrm{Var}(X) = np(1-p), \qquad \mathrm{SD}(X) = \sqrt{np(1-p)}.$$

For the 20-image batch, $\mathbb{E}[X] = 20(0.865) = 17.3$ correct and $\mathrm{SD}(X) = \sqrt{20(0.865)(0.135)} \approx 1.53$. Most batches therefore land within a couple of predictions of 17.

EXAMPLE 5.6 • At least one failure

Each of $n = 10$ independent requests fails with probability $p = 0.02$. The probability of *at least one* failure is awkward to sum directly, but its complement — no failures at all — is a single binomial term:

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - (0.98)^{10} \approx 0.183.$$

Even a 2% per-request failure rate yields a roughly 18% chance that something fails somewhere in the batch — the reasoning behind retries and redundancy.

5.3.4 Normal approximation to the binomial

As the expected counts np and $n(1-p)$ both grow, the binomial pmf becomes more symmetric and bell-shaped (Figure 5.7). When it is symmetric enough, the normal distribution with the same mean and standard deviation is an excellent approximation, which spares us from summing many terms.

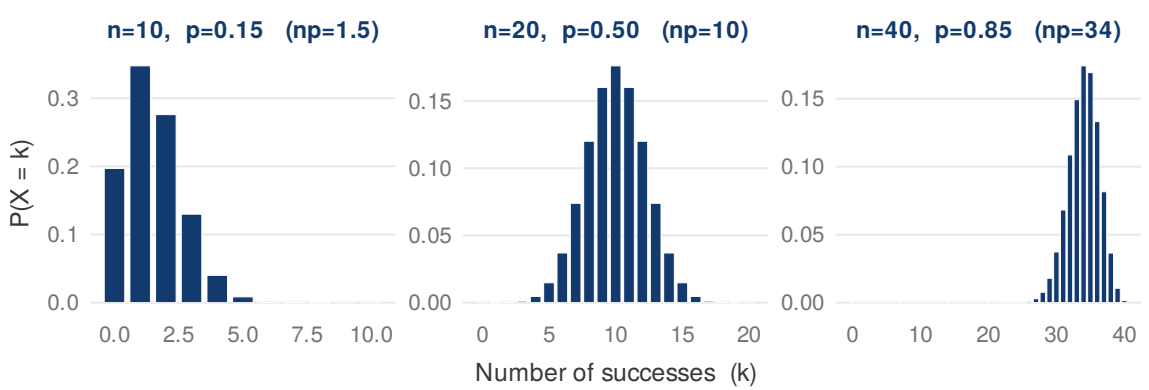


Figure 5.7. Binomial shape depends on *both* np and $n(1-p)$. When np is small the distribution is right-skewed (left); when np and $n(1-p)$ are both large and comparable it is approximately symmetric (center); when $n(1-p)$ is small the distribution is left-skewed (right, where $np = 34$ but $n(1-p) = 6$). Symmetry requires both the expected successes np and the expected failures $n(1-p)$ to be large.

NORMAL APPROXIMATION AND THE SUCCESS–FAILURE CONDITION

If $np \geq 10$ *and* $n(1-p) \geq 10$ (the **success–failure condition**), then $X \sim B(n,p)$ is well approximated by the normal distribution

$$X \;\dot{\sim}\; N\big(np,\; np(1-p)\big).$$

The symbol $\dot{\sim}$ — a dot placed over the tilde — is read "is approximately distributed as," in contrast to the plain $\sim$, which asserts an *exact* distribution. We use it whenever a distribution is being approximated by another.

EXAMPLE 5.7 • Scaling up the evaluation

Evaluate the same classifier on $n = 100$ images. Then $np = 86.5 \geq 10$ and $n(1-p) = 13.5 \geq 10$, so the success–failure condition holds and the number correct is approximately

$$X \;\dot{\sim}\; N\big(86.5,\; 3.42^2\big), \qquad \mathrm{SD}(X) = \sqrt{100(0.865)(0.135)} \approx 3.42.$$

Figure 5.8 shows the close agreement between the exact bars and the approximating curve.

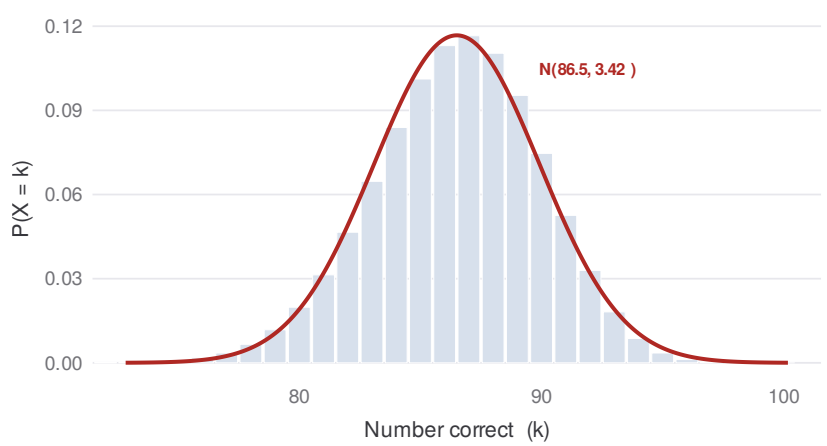


Figure 5.8. Binomial pmf for $B(100, 0.865)$ (bars) with its normal approximation $N(86.5, 3.42^2)$ (curve). The fit is excellent because the success–failure condition is met.

! CAUTION

The approximation degrades when n is small or p is near 0 or 1, exactly when the success–failure condition fails and the binomial is still noticeably skewed. In those cases compute binomial probabilities exactly.

EXERCISES (§5.3)

1. Suppose that 60% of students at a university hold a driver's license. 10 students are randomly selected, and X denotes the number of students among them who hold a driver's license. Answer the following questions.
   (a) Find the probability that exactly 8 or exactly 9 of the selected students hold a driver's license.
   (b) Find the expectation and variance of the random variable X.

2. A biased coin is tossed $n = 3$ times, where the probability of heads is $p = \frac{1}{3}$. Let X be the number of heads and Y be the number of tails. Answer the following questions.
   (a) Find the probability mass function $f_X(x)$ of X.
   (b) The joint pmf of X and Y is given in the table below. Find the values (1) and (2).

$x \setminus y$	0	1	2	3
0	0	0	0	(1)
1	0	0	12/27	0
2	0	6/27	0	0
3	1/27	(2)	0	0

   (c) Determine whether X and Y are independent.(Hint: the probability mass function of Y is $f_Y(y) = \binom{3}{y}\left(\frac{2}{3}\right)^y\left(\frac{1}{3}\right)^{3-y}$, $\quad y = 0, 1, 2, 3.$)
   (d) Find $\mathrm{Var}(X - Y)$.

3. Boston Children's Hospital estimates that 90% of Americans have had chickenpox by the time they reach adulthood.
   (a) Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.
   (b) What is the probability that at least 1 out of 10 randomly sampled American adults has had chickenpox?

4. Boston Children's Hospital reports that $p = 90\%$ of American adults had chickenpox during childhood. Given that $n = 144$ American adults are randomly selected, let the random variable X denote the number of individuals who had chickenpox during childhood. Answer the following.
   (a) Calculate $\mathbb{P}(X = 2) \times 10^{142}$.
   (b) Find $E(X)$ and $\mathrm{Var}(X)$.
   (c) Find the probability that 135 or more individuals had chickenpox during childhood. Assume that the conditions for the Central Limit Theorem are satisfied.

5. A factory produces batteries with a known defect rate of 0.1. Ten batteries are randomly selected and sent to Facility P.
   (a) Let the random variable X denote the number of defective batteries among the ten sent to Facility P. Determine the probability distribution of X, and find $E(X)$ and $\mathrm{Var}(X)$.
   (b) Let $Y = \frac{X}{10}$. Find $E(Y)$ and $\mathrm{Var}(Y)$.
   (c) Calculate the probability that at least two defective batteries are sent to Facility P.

CHAPTER 6

Foundations for Inference

The previous chapters taught us to summarize data (Chapter 2) and to model randomness with distributions (Chapter 5). We now begin **statistical inference**: using a sample to draw conclusions about the larger population it came from. Three tasks recur throughout. A **point estimate** gives a single best guess for an unknown population quantity, a **confidence interval** surrounds that guess with an honest range of uncertainty, and a **hypothesis test** weighs the evidence for a specific claim. This chapter develops the first two for a population **proportion** p — the natural target whenever each observation is a success or a failure — with the Central Limit Theorem as the engine that makes them work. The third, hypothesis testing, is taken up in Chapter 8, where the same reasoning is developed in full for a population mean.

6.1 Point estimates and sampling variability

Suppose we want to know the true accuracy p of a trained classifier — the proportion of *all* possible inputs it would label correctly. We cannot test every possible input, so we evaluate the model on a finite test set of n examples and record the number correct, X. As in Chapter 5, $X \sim B(n, p)$. The

sample proportion

$$\hat{p} = \frac{X}{n} = \frac{\text{number correct}}{\text{number tested}}$$

is our **point estimate** of p.

DEFINITION 6.1 • Point estimate

A **point estimate** is a single number, computed from a sample, used as a best guess for an unknown **population parameter**. The sample proportion $\hat{p}$ is the point estimate of the population proportion p; the hat (ˆ) denotes "estimated from data."

6.1.1 Sampling variability and the standard error

A point estimate is almost never exactly right. Had we drawn a different test set we would have computed a different $\hat{p}$; the gap $\hat{p} - p$ is the **sampling error**. To reason about how large that error tends to be, imagine repeating the whole experiment many times — a fresh sample of size n each time — and collecting all the resulting $\hat{p}$ values. Their distribution is the **sampling distribution** of $\hat{p}$ (Figure 6.1).

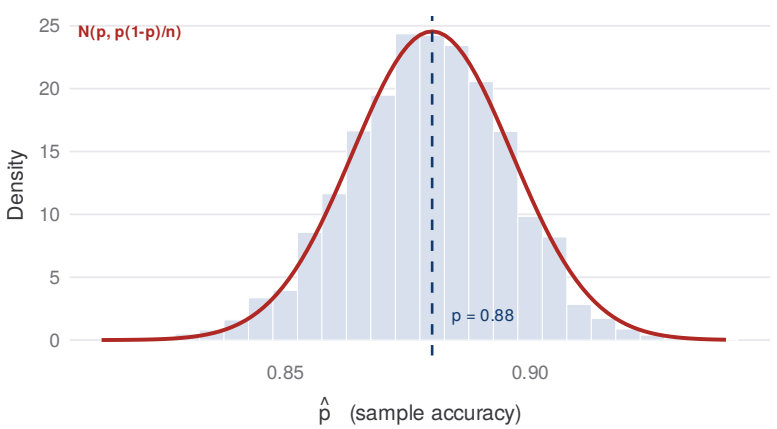


Figure 6.1. Sampling distribution of $\hat{p}$ from 5,000 simulated test sets of size $n = 400$ when the true accuracy is $p = 0.88$. It is centered at p, bell-shaped, and closely matched by a normal curve (red).

! CAUTION

In practice we collect only *one* sample, so the sampling distribution is never actually observed. It is a thought experiment — but a powerful one, because its center and spread tell us how much to trust the single estimate we do have.

The spread of the sampling distribution has a special name.

DEFINITION 6.2 • Standard error

The **standard error** of a point estimate is the standard deviation of its sampling distribution. For the sample proportion,

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

This formula follows directly from Chapter 5. Since $\hat{p} = X/n$ with $X \sim B(n, p)$, and $\text{Var}(X) = np(1-p)$,

$$\mathbb{E}[\hat{p}] = \frac{\mathbb{E}[X]}{n} = p, \qquad \text{Var}(\hat{p}) = \frac{\text{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

The standard error is the square root of that variance. Because n sits in the denominator, $SE(\hat{p})$ shrinks as the sample grows: larger test sets yield more consistent estimates (Figure 6.2).

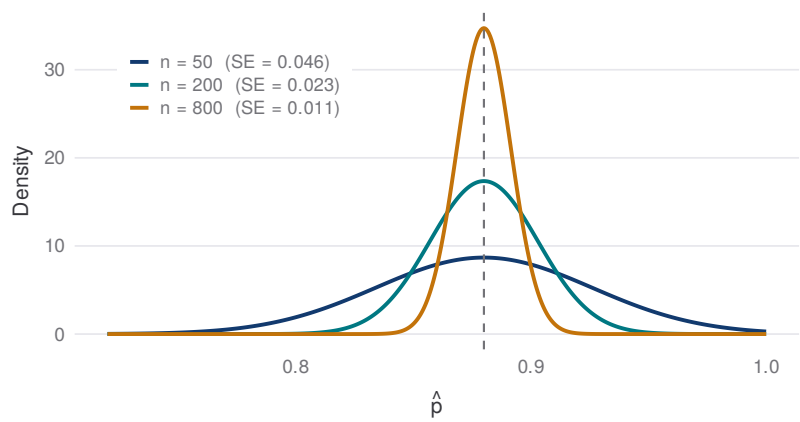


Figure 6.2. The sampling distribution of $\hat{p}$ narrows as n increases (here $p = 0.88$). Quadrupling n halves the standard error, since $SE \propto 1/\sqrt{n}$.

6.1.2 The Central Limit Theorem for a proportion

The simulated sampling distribution in Figure 6.1 looks normal for a reason. The **Central Limit Theorem** (CLT) guarantees it, provided two conditions hold.

CENTRAL LIMIT THEOREM FOR THE SAMPLE PROPORTION

When (1) the observations are **independent** and (2) the **success–failure condition** $np \geq 10$ and $n(1-p) \geq 10$ is met, the sampling distribution of $\hat{p}$ is approximately normal:

$$\hat{p} \;\dot{\sim}\; N\left(p,\ \frac{p(1-p)}{n}\right).$$

NOTE

Condition (1) holds automatically when the sample is drawn *with* replacement. When sampling *without* replacement from a finite population — the usual case — the observations are not strictly independent, but the dependence is negligible as long as the sample is a small fraction of the population: the **10% condition**, $n \le 0.10\,N$, where N is the population size. For larger sampling fractions a finite-population correction would be needed.

Recall from Chapter 5 that $\dot{\sim}$ reads "is approximately distributed as." The approximation improves as n grows and is excellent once the success–failure condition is comfortably satisfied.

! CAUTION

The formula for $SE(\hat{p})$ contains the unknown p. In practice we substitute the estimate $\hat{p}$:

$$SE(\hat{p}) \approx \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

We likewise check the success–failure condition with $\hat{p}$ in place of p.

EXAMPLE 6.1 • Estimating a classifier's accuracy

A model is evaluated on $n = 400$ independent test images and labels 352 of them correctly, so

$$\hat{p} = \frac{352}{400} = 0.88.$$

The successes (352) and failures (48) both exceed 10, and the images are independent, so the CLT applies:

$$\hat{p} \;\dot{\sim}\; N\!\left(p,\ \frac{p(1-p)}{n}\right),$$

centered at the *unknown* true accuracy p. The standard error $\sqrt{p(1-p)/n}$ is itself unknown, so we estimate it by plugging in $\hat{p} = 0.88$:

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{0.88(1-0.88)}{400}} = \sqrt{\frac{0.1056}{400}} \approx 0.016.$$

Across repeated test sets of this size, the estimated accuracy would vary about p with a standard deviation of about 1.6 percentage points.

GUIDED PRACTICE 6.1

A spam filter is checked on 50 messages and flags the correct label on 48. A colleague wants to invoke the CLT for $\hat{p}$. Is the success–failure condition met?
Answer: the failures number only $50 - 48 = 2 < 10$, so the condition fails and the normal

approximation should not be trusted here; an exact binomial calculation is safer.

EXERCISES (§6.1)

1. It is known that 20% of students at a university live in the campus dormitory. A random sample of $n = 60$ students is selected, and X students among them live in the dormitory. Answer the following questions.
   (a) Find the approximate distribution of the sample proportion $\hat{p} = X/n$.
   (b) Find $\mathbb{P}\left(\hat{p} \geq \frac{1}{6}\right)$.
2. Suppose the proportion of all Korean adults who support the expansion of renewable energy is $p = 0.75$. We randomly sample $n = 300$ Korean adults and consider the sample proportion $\hat{p} = X/300$, where X is the number of people who support the expansion of renewable energy out of the 300 people.
   (a) Find the mean and standard error of $\hat{p}$.
   (b) Find the probability that $\hat{p}$ is less than 0.74, $\mathbb{P}(\hat{p} < 0.74)$. (Suppose that the normality condition is met.)
3. A factory produces aircraft parts with a known defect rate of $p = 0.02$. Let the random variable X denote the number of defective aircraft parts in a random sample of $n = 1000$ parts. Answer the following.
   (a) Calculate $\mathbb{P}(X = 0) \times 10^9$.
   (b) Let $\hat{p} = \frac{X}{1000}$ denote the sample proportion of defective parts. Find the expectation and standard error of $\hat{p}$, $E(\hat{p})$ and $SE(\hat{p})$.
   (c) Find $\mathbb{P}(\hat{p} < 0.019)$ using the normal approximation of $\hat{p}$. (Check the normality condition if necessary.)

6.2 Confidence intervals for a proportion

A point estimate answers "what is our best single guess?" but says nothing about how far off it might be. A **confidence interval** answers the better question: "what range of values for p is plausible, given the data?" Reporting $\hat{p} = 0.88$ alone is a spear; reporting an interval is a net.

DEFINITION 6.3 • Confidence interval

A **confidence interval** is a range of plausible values for a population parameter, computed from a sample so that, over many repeated samples, a fixed percentage of such intervals contain the true parameter. That percentage is the **confidence level**.

6.2.1 Constructing a 95% interval

When the point estimate is approximately normal, the empirical rule of Chapter 5 does the work: about 95% of the time a normal estimate lands within (almost exactly) 1.96 standard errors of its target. Turning that around gives an interval around the estimate.

95% CONFIDENCE INTERVAL FOR A PROPORTION

When the CLT conditions of Section 6.1 hold, an approximate 95% confidence interval for the population proportion p is

$$\hat{p} \pm 1.96 \times SE(\hat{p}), \qquad SE(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}.$$

EXAMPLE 6.2 • A 95% interval for the model's accuracy

Continuing Example 6.1, we had $\hat{p} = 0.88$ and $SE(\hat{p}) \approx 0.016$ from $n = 400$ images. The 95% confidence interval is

$$0.88 \pm 1.96(0.016) = 0.88 \pm 0.031 = (0.849,\ 0.911).$$

We are 95% confident that the model's true accuracy lies between about 84.9% and 91.1%. Notice the interval is a statement about the unknown p, expressed with the modesty that a finite test set demands.

6.2.2 What "95% confident" means

The confidence level describes the long-run behavior of the *procedure*, not a single interval. If we drew many samples and built an interval from each, about 95% of those intervals would capture the true p (Figure 6.3).

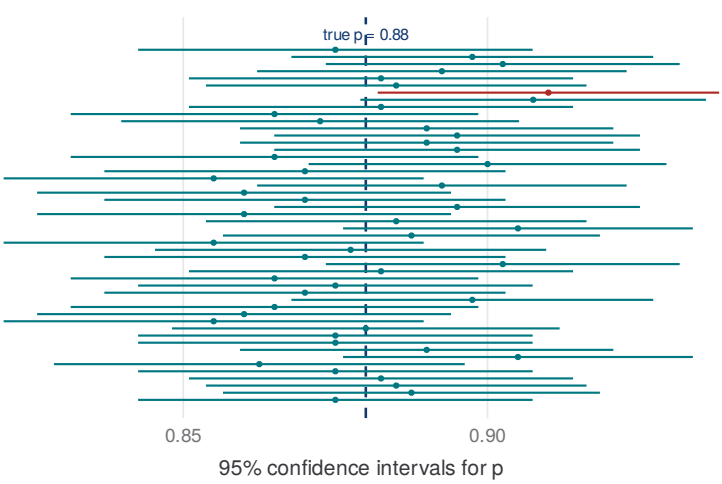


Figure 6.3. Fifty 95% confidence intervals, each from a fresh simulated sample ($p = 0.88$). Most capture the true value (dashed line); the rare interval that misses it is drawn in red. In the long run about 5% miss.

! CAUTION

It is tempting but wrong to say "there is a 95% probability that p is in $(0.849, 0.911)$." The parameter p is a fixed (if unknown) number, and this particular interval either contains it or does not. The 95% refers to how often the *method* succeeds across repetitions.

6.2.3 Margin of error and other confidence levels

The quantity added and subtracted is the **margin of error**. For a 95% interval it is $1.96 \times SE$; the full interval width is twice that. To change the confidence level we simply change the multiplier.

CONFIDENCE INTERVAL AT ANY LEVEL

If a point estimate is approximately normal, a $100(1-\alpha)\%$ confidence interval for the parameter is

$$\text{point estimate} \pm \underbrace{z_{\alpha/2} \times SE}_{\text{margin of error}},$$

where $z_{\alpha/2}$ is the standard-normal cutoff leaving area $\alpha/2$ in the upper tail. Common values are

$$z_{0.05} = 1.645 \ (90\%), \quad z_{0.025} = 1.96 \ (95\%), \quad z_{0.005} = 2.576 \ (99\%).$$

A higher confidence level requires a larger multiplier and therefore a *wider* interval (Figure 6.4): being more certain of capturing p means casting a bigger net. The only way to be both more confident *and* more precise is to collect more data, which shrinks SE.

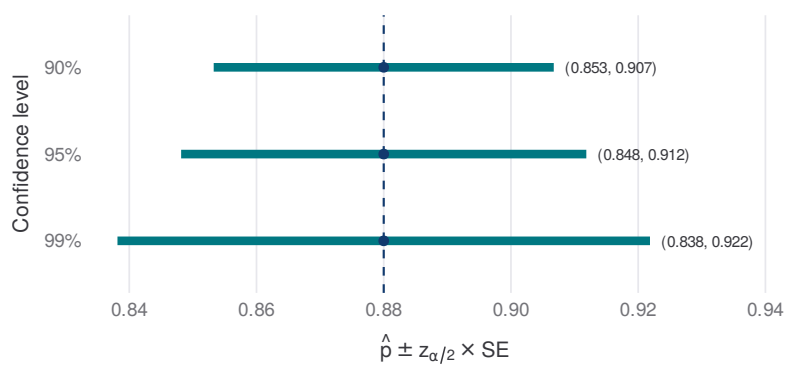


Figure 6.4. Confidence intervals for the same estimate ($\hat{p} = 0.88$, $n = 400$) at three levels. Higher confidence widens the interval because $z_{\alpha/2}$ grows.

GUIDED PRACTICE 6.2

Using $\hat{p} = 0.88$ and $SE(\hat{p}) = 0.016$, construct a 90% confidence interval for the model's

accuracy.
Answer: $0.88 \pm 1.645(0.016) = 0.88 \pm 0.026 = (0.854, 0.906)$ — narrower than the 95% interval, as expected.

EXERCISES (§6.2)

1. In a survey, 1,000 randomly selected people were asked whether they believe it is good to consider blood type when making friends; 214 responded that they believe it is.
   (a) Check whether the conditions required for constructing a confidence interval for a population proportion are satisfied.
   (b) Construct a 95% confidence interval for the proportion of people who believe it is good to consider blood type when making friends.
2. A city tested a beta version of a new public transportation app by surveying a random sample of $n = 850$ commuters. Among them, $x = 137$ reported being "very satisfied." Answer the following.
   (a) Construct an approximate 95% confidence interval for the proportion p of users who reported being "very satisfied." (You do not need to check the conditions for the Central Limit Theorem.)
   (b) Choose the most appropriate word to complete each sentence.
      i. As the sample size increases, the standard error of the sample proportion $\hat{p}$ becomes (larger / smaller).
      ii. As the sample size increases, the length of the confidence interval becomes (larger / smaller).
      iii. The length of a 99% confidence interval is (larger / smaller) than that of a 95% confidence interval.
3. In a university survey of 800 students, 440 reported that they smoke. Construct a 95% approximate confidence interval for the population proportion of students who smoke. (Assume that the conditions for the Central Limit Theorem are satisfied.) Use $z_{0.025} = 1.96$.

CHAPTER 7

Sampling Distribution and Point Estimation

Chapter 6 introduced the first tools of inference — a point estimate and a confidence interval — for a single population proportion, resting on the sampling distribution of $\hat{p}$. The remarkable fact is that the same ideas apply to *any* sample statistic. This chapter makes that generality explicit. Section 7.1 introduces the **sampling distribution** of a statistic and works out three central cases — the sample mean (with the Central Limit Theorem), the sample variance (with the chi-squared distribution), and the sample proportion. Section 7.2 then asks what makes a good estimator, formalizing **unbiasedness**, **efficiency**, and **consistency**. Together they are the foundation for the inference about means that follows.

7.1 Sampling distribution

7.1.1 Population, sample, and statistic

Statistical inference begins by distinguishing the whole from the part.

DEFINITION 7.1 • Population, parameter, and sample

The **population** is the entire set of units under study. A **parameter** is a numerical feature of the population, such as its mean μ, variance σ^2, or a proportion p. A **sample** is a subset drawn from the population. Because measuring the entire population is usually impossible, we use a sample to estimate parameters.

We model the sample as random.

DEFINITION 7.2 • Random sample

A **random sample** of size n from a population with distribution F is a collection of n independent random variables

$$X_1, X_2, \dots, X_n \sim F,$$

each with the same distribution as the population. Such variables are **independent and identically distributed** (i.i.d.).

Any quantity computed from the sample is itself random, and so has a distribution of its own.

DEFINITION 7.3 • Statistic and sampling distribution

A **statistic** is a function of the random sample that does not involve unknown parameters. Since it is computed from random data, a statistic is itself a random variable, and its probability distribution is called its **sampling distribution**. Two basic statistics are the **sample mean** and **sample variance**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

EXAMPLE 7.1 • A sampling distribution by hand

Suppose the number of retries a request needs, X, takes the values $0, 1, 2$ with probabilities 0.3, 0.6, 0.1. Draw a random sample of size $n = 2$ and consider the sample mean $\overline{X} = (X_1 + X_2)/2$. Because X_1 and X_2 are independent, each cell of their joint distribution is a product of marginals, e.g. $\mathbb{P}(X_1 = 0, X_2 = 1) = 0.3 \times 0.6 = 0.18$.

$X_1 \backslash X_2$	0	1	2
0	0.09	0.18	0.03
1	0.18	0.36	0.06
2	0.03	0.06	0.01

Collecting the cells by the value of $\overline{X}$ gives its sampling distribution:

$\overline{X}$	0	0.5	1	1.5	2
$\mathbb{P}(\overline{X})$	0.09	0.36	0.42	0.12	0.01

As a check, the population mean is $\mu = 0(0.3) + 1(0.6) + 2(0.1) = 0.8$, and the mean of this sampling distribution is likewise $0(0.09) + 0.5(0.36) + 1(0.42) + 1.5(0.12) + 2(0.01) = 0.8$.

GUIDED PRACTICE 7.1

For the same retry variable ($X \in \{0, 1, 2\}$ with probabilities $0.3, 0.6, 0.1$) and a sample of size $n = 2$, find the sampling distribution of the sample *maximum* $M = \max(X_1, X_2)$.
Answer: reading the joint table by the larger coordinate, $\mathbb{P}(M = 0) = 0.09$, $\mathbb{P}(M = 1) = 0.72$, and $\mathbb{P}(M = 2) = 0.19$ (which sum to 1). The maximum has a different sampling distribution from the mean — every statistic carries its own.

7.1.2 The sample mean

The example's check — that $\overline{X}$ is centered at μ — is no accident. It holds for any population.

MEAN AND VARIANCE OF THE SAMPLE MEAN

For a random sample of size n from a population with mean μ and variance σ^2,

$$\mathbb{E}[\overline{X}] = \mu, \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}, \qquad SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Both follow from the rules for sums of independent variables: by linearity, $\mathbb{E}[\overline{X}] = \frac{1}{n}\sum \mathbb{E}[X_i] = \frac{1}{n}\, n\mu = \mu$, and because the X_i are independent, $\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum \mathrm{Var}(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}$. The standard error therefore shrinks as the sample grows.

When the population is itself normal, the sample mean is exactly normal, since a linear combination of independent normals is normal.

SAMPLE MEAN OF A NORMAL POPULATION

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, then

$$\overline{X} \sim N\!\left(\mu,\ \frac{\sigma^2}{n}\right), \qquad Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

EXAMPLE 7.2 • Average inference time

A model's per-image inference time is $X \sim N(20, 2^2)$ ms. For a batch of $n = 4$ images, the average $\overline{X}$ satisfies

$$\overline{X} \sim N\left(20, \frac{2^2}{4}\right) = N(20, 1),$$

so

$$\mathbb{P}(\overline{X} \le 19) = \mathbb{P}\left(Z \le \frac{19 - 20}{1}\right) = \mathbb{P}(Z \le -1) \approx 0.159.$$

GUIDED PRACTICE 7.2

For the same model, $X \sim N(20, 2^2)$ ms, take a larger batch of $n = 16$ images. Find $\mathbb{P}(\overline{X} \le 19)$ and compare it with the $n = 4$ result above.
Answer: now $\overline{X} \sim N(20,\ 2^2/16) = N(20,\ 0.5^2)$, so $\mathbb{P}(\overline{X} \le 19) = \mathbb{P}(Z \le (19 - 20)/0.5) = \mathbb{P}(Z \le -2) \approx 0.023$ — well below the 0.159 at $n = 4$, because averaging more images concentrates $\overline{X}$ near μ.

7.1.3 The Central Limit Theorem

If the population is not normal — as inference latency, heavily right-skewed in Chapter 2, is not — the sample mean is still approximately normal once n is large.

CENTRAL LIMIT THEOREM (CLT)

For a random sample $X_1, \ldots, X_n$ from any population with mean μ and finite variance σ^2, when n is large

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \ \dot{\sim}\ N(0, 1), \qquad \text{equivalently} \qquad \overline{X} \ \dot{\sim}\ N\left(\mu, \frac{\sigma^2}{n}\right),$$

regardless of the shape of the population distribution.

Figure 7.1 shows the effect for the skewed latency population: the distribution of $\overline{X}$ is already roughly symmetric by $n = 30$ and indistinguishable from normal by $n = 100$, all the while tightening around μ.

! CAUTION

The CLT describes the *mean*, not individual observations: with $n = 100$ the average latency is nearly normal, but a single latency is still as skewed as ever. How large n must be depends on the population — mildly skewed data need only a modest sample, while heavily skewed data need more.

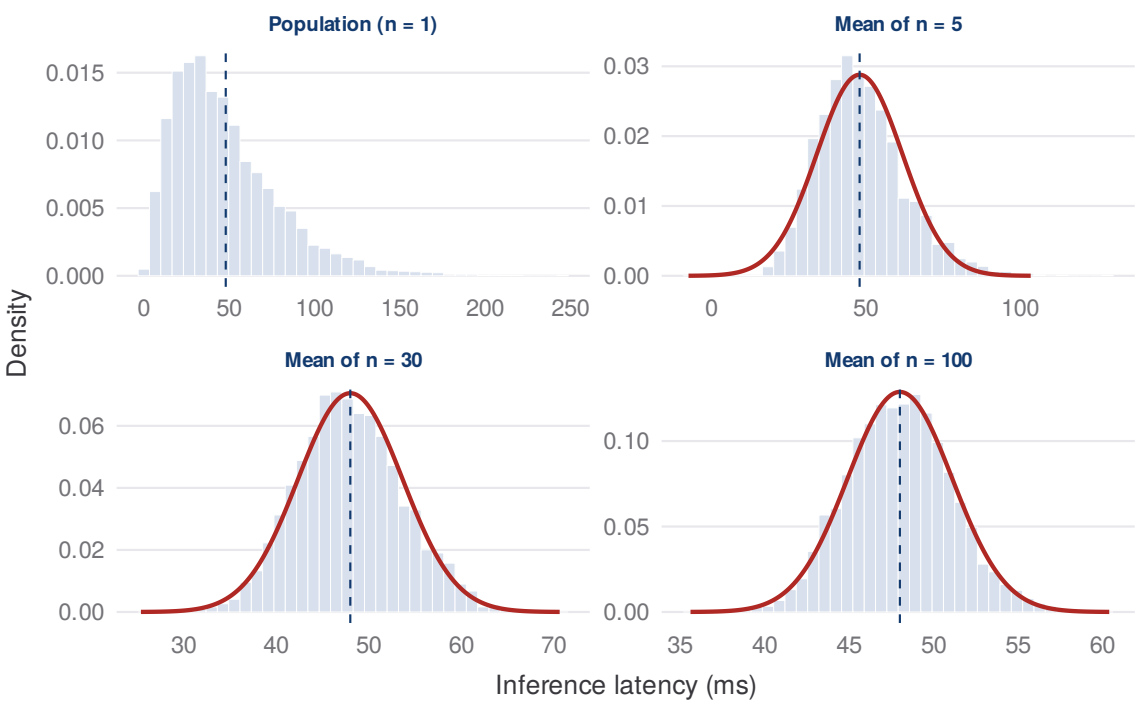


Figure 7.1. The Central Limit Theorem at work. The population of latencies (top left) is right-skewed, but the sampling distribution of $\overline{X}$ approaches the normal curve (red) and narrows as n grows from 5 to 100.

EXAMPLE 7.3 • Average response length

The number of tokens in a model's response has mean $\mu = 266$ and standard deviation $\sigma = 15$. For a random sample of $n = 81$ responses, n is large enough to apply the CLT:

$$\overline{X} \dot{\sim} N\left(266, \left(\tfrac{15}{9}\right)^2\right) = N\left(266, \left(\tfrac{5}{3}\right)^2\right),$$

so

$$\mathbb{P}(\overline{X} \geq 270) = \mathbb{P}\left(Z \geq \frac{270 - 266}{5/3}\right) = \mathbb{P}(Z \geq 2.4) \approx 0.008.$$

GUIDED PRACTICE 7.3

A job queue's service time has mean $\mu = 4.0$ s and standard deviation $\sigma = 2.5$ s, with a right-skewed shape. For a random sample of $n = 64$ jobs, approximate $\mathbb{P}(\overline{X} < 3.6)$.
Answer: n is large, so the CLT gives $\overline{X} \dot{\sim} N\left(4.0,\ (2.5/8)^2\right)$ with $SE = 0.3125$. Then $\mathbb{P}(\overline{X} < 3.6) = \mathbb{P}\left(Z < (3.6 - 4.0)/0.3125\right) = \mathbb{P}(Z < -1.28) \approx 0.10$. The skew of a single service time does not matter here — only the mean is being averaged.

7.1.4 The chi-squared distribution and the sample variance

To make inference about a population *variance* σ^2 we use the sample variance S^2, whose distribution is governed by a new family: the chi-squared distribution.

DEFINITION 7.4 • Chi-squared distribution

If $Z_1, Z_2, \ldots, Z_k \sim N(0, 1)$ independently, then the sum of their squares

$$V = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

follows the **chi-squared distribution** with k **degrees of freedom**, written $V \sim \chi^2(k)$.

REMARK 7.1 • Shape of the chi-squared distribution

A $\chi^2(k)$ variable has mean k and variance $2k$, so as the degrees of freedom increase, its center and spread both grow (Figure 7.2). It also becomes *less* skewed, approaching a normal shape — not more skewed, a common misconception.

TWO PROPERTIES OF THE CHI-SQUARED DISTRIBUTION

Additivity: if $V_1 \sim \chi^2(k_1)$ and $V_2 \sim \chi^2(k_2)$ independently, then $V_1 + V_2 \sim \chi^2(k_1 + k_2)$.

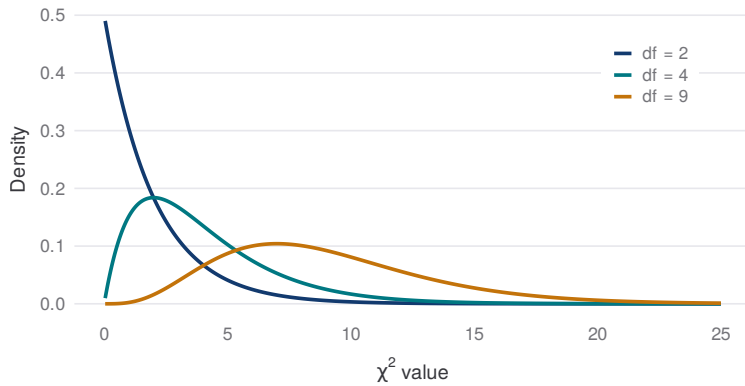


Figure 7.2. Chi-squared densities for several degrees of freedom. For small k the distribution is sharply right-skewed; as k grows it shifts right, spreads out, and becomes more symmetric.

Square of a standard normal: if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$.

These properties yield the sampling distribution of S^2 for a normal population.

DISTRIBUTION OF THE SAMPLE VARIANCE

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ independently. Then $\overline{X}$ and S^2 are independent, and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

NOTE

This result is the foundation for confidence intervals and hypothesis tests about a variance σ^2.

EXAMPLE 7.4 • A probability for the sample variance

A sensor's reading noise is modeled as $N(\mu, \sigma^2)$ with a known $\sigma^2 = 4$. From $n = 10$ readings, how often would the sample variance exceed 8? Scaling by the result above turns the question about S^2 into one about a $\chi^2(9)$ variable (Figure 7.2):

$$\mathbb{P}(S^2 > 8) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma^2} > \frac{9 \times 8}{4}\right) = P\big(\chi^2(9) > 18\big) \approx 0.035.$$

So even with the true variance fixed at 4, a sample variance above 8 arises about 3.5% of the

time — sampling alone makes S^2 swing widely when n is small.

```
> pchisq(18, df = 9, lower.tail = FALSE)
[1] 0.0351738
```

NOTE

Turning this around — solving for the cutoff that leaves a fixed tail area — is exactly how a confidence interval or test for a variance σ^2 is built, the variance analogue of the mean procedures in Chapter 8.

7.1.5 The sample proportion

When each observation is a success or failure, the population is Bernoulli and the parameter of interest is the proportion p.

For $X_i \sim B(1, p)$ drawn independently, the **sample proportion** is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since the number of successes $\sum_{i=1}^{n} X_i \sim B(n, p)$, dividing by n gives

$$\mathbb{E}(\hat{p}) = p, \qquad \operatorname{Var}(\hat{p}) = \frac{p(1-p)}{n},$$

recovering the standard error of Chapter 6. The CLT then applies to this average just as it does to any other.

CLT FOR THE SAMPLE PROPORTION

When n is large (in practice $np \geq 10$ and $n(1-p) \geq 10$),

$$\hat{p} \;\dot{\sim}\; N\!\left(p, \frac{p(1-p)}{n}\right), \qquad \text{equivalently} \qquad \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \;\dot{\sim}\; N(0,1).$$

EXAMPLE 7.5 • A model's success rate in a sample

A model answers each query correctly with probability $p = 0.6$. In a random sample of $n = 30$ queries, what is the chance the sample success rate $\hat{p}$ falls below 0.5? Here

$$\mathbb{E}(\hat{p}) = 0.6, \qquad \operatorname{Var}(\hat{p}) = \frac{0.6 \times 0.4}{30} = 0.008,$$

and since $np = 18 \geq 10$ and $n(1-p) = 12 \geq 10$, the CLT gives $\hat{p} \dot{\sim} N(0.6, 0.008)$. Thus

$$\mathbb{P}(\hat{p} \leq 0.5) = \mathbb{P}\left(Z \leq \frac{0.5 - 0.6}{\sqrt{0.008}}\right) = \mathbb{P}(Z \leq -1.12) \approx 0.131.$$

GUIDED PRACTICE 7.4

A deployed classifier is correct on a fraction $p = 0.8$ of inputs. In a random audit of $n = 100$ inputs, what is the approximate probability the sample accuracy $\hat{p}$ exceeds 0.85?
Answer: $\mathbb{E}(\hat{p}) = 0.8$ and $\text{Var}(\hat{p}) = 0.8(0.2)/100 = 0.0016$, a standard deviation of 0.04; with $np = 80$ and $n(1-p) = 20$ both at least 10, the CLT gives $\hat{p} \dot{\sim} N(0.8,\ 0.04^2)$. Then $\mathbb{P}(\hat{p} > 0.85) = \mathbb{P}(Z > (0.85 - 0.80)/0.04) = \mathbb{P}(Z > 1.25) \approx 0.106$.

EXERCISES (§7.1)

1. The lifetime of portable flashlights produced at Factory K has a mean of 1,500 hours and a standard deviation of 100 hours. If 100 flashlights are randomly selected from this factory, find the probability that the sample mean lifetime $\overline{X}$ exceeds 1,510 hours.

2. In a certain sport, Team A competes against Team B, with probabilities of winning and losing recorded as 0.6 and 0.4, respectively. A win grants 3 points, while a loss grants 0 points. If Team A plays n matches, and the points scored in each match are denoted as $X_1, X_2, \ldots, X_n$, answer the following questions. (Assume no draws, identical conditions for all matches, and independent results.)

   (a) Find the probability mass function (PMF) of the average points scored in two matches, $\overline{X}_2$.

   (b) Compute the expected value $E(\overline{X}_2)$ and variance $Var(\overline{X}_2)$.

   (c) Find the approximate probability that the average score of Team A is at least 2 points in 54 matches.

3. The monthly income of residents in a certain region follows a normal distribution $N(200, 200)$ with mean $\mu = 200$ and variance $\sigma^2 = 200$. Two residents are randomly selected, and their monthly incomes X_1 and X_2 are observed. (Assume that X_1 and X_2 are independent.) Answer the following.

   (a) Find the expectation and variance of the average income $\overline{X} = \dfrac{X_1 + X_2}{2}$, $E(\overline{X})$ and $\text{Var}(\overline{X})$.

   (b) Compute the probability that the average income $\overline{X}$ is less than 180, $\mathbb{P}(\overline{X} < 180)$.

4. From a population having the following probability distribution, a random sample X_1 and X_2 of size $n = 2$ is taken.

x	1	2	3
$\mathbb{P}(X = x)$	0.6	0.2	0.2

The joint probability distribution of X_1 and X_2, $\mathbb{P}(X_1 = x_1, X_2 = x_2)$, is given below.

$x_1 \backslash x_2$	1	2	3
1	0.36	0.12	0.12
2	0.12	0.04	0.04
3	0.12	0.04	0.04

   (a) Find the probability distribution of the sample mean $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

   (b) For the sample variance $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, show that $S^2 = \dfrac{1}{2}(X_1 - X_2)^2$.

(c) The probability distribution of the sample variance S^2 is given below. Determine whether the sample mean $\overline{X}$ and the sample variance S^2 are independent.

s^2	0	0.5	2	Otherwise
$\mathbb{P}(S^2 = s^2)$	0.44	0.32	0.24	0

7.2 Point estimation

7.2.1 Estimators and estimates

Point estimation uses sample information to produce a single best guess for an unknown parameter. It is worth distinguishing the rule from the number it produces.

DEFINITION 7.5 • Estimator and estimate

An **estimator** $\hat{\theta}(X_1, \ldots, X_n)$ is a statistic used to estimate a parameter θ; as a function of random data it is a random variable. The **estimate** $\hat{\theta}(x_1, \ldots, x_n)$ is the number obtained by plugging in observed data. We write a generic estimator of θ as $\hat{\theta}$.

	Mean	**Variance**
Parameter	μ	σ^2
Estimator	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
Estimate	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

7.2.2 Properties of estimators

Several estimators can target the same parameter, so we judge them by the behavior of their sampling distributions. Three properties are desirable: **unbiasedness**, **efficiency**, and **consistency** (Figure 7.3).

DEFINITION 7.6 • Unbiasedness

An estimator $\hat{\theta}$ of θ is **unbiased** if

$$\mathbb{E}(\hat{\theta}) = \theta.$$

Its **bias** is $B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$, the systematic error remaining after averaging over all samples.

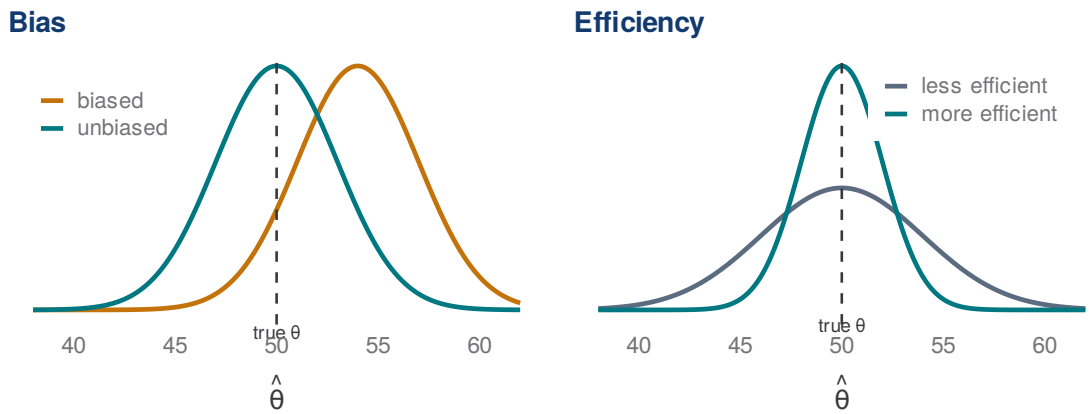


Figure 7.3. Left: an *unbiased* estimator is centered on the true θ; a *biased* one is systematically off. Right: among unbiased estimators, the *more efficient* one has the smaller variance and so is more reliable.

DEFINITION 7.7 • Efficiency

Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of the same parameter, $\hat{\theta}_1$ is **more efficient** if it has the smaller variance,

$$\operatorname{Var}(\hat{\theta}_1) < \operatorname{Var}(\hat{\theta}_2).$$

EXAMPLE 7.6 • Comparing two estimators of the mean

Let X_1, X_2, X_3 be a random sample with mean μ and variance σ^2, and consider

$$T_1 = \frac{X_1 + 2X_2 + X_3}{4}, \qquad T_2 = \frac{X_1 + X_2 + X_3}{3}.$$

Both are unbiased ($\mathbb{E}[T_1] = \mathbb{E}[T_2] = \mu$). Their variances are

$$\operatorname{Var}(T_1) = \frac{1^2 + 2^2 + 1^2}{4^2}\,\sigma^2 = \frac{3}{8}\sigma^2, \qquad \operatorname{Var}(T_2) = \frac{3}{3^2}\,\sigma^2 = \frac{1}{3}\sigma^2.$$

Since $\frac{3}{8}\sigma^2 > \frac{1}{3}\sigma^2$, the equally weighted average T_2 is more efficient.

DEFINITION 7.8 • Consistency

An estimator $\hat{\theta}_n$ of θ (based on a sample of size n) is **consistent** if it converges to θ *in probability*: for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|\hat{\theta}_n - \theta| > \varepsilon\big) = 0.$$

A convenient sufficient condition is that the estimator be asymptotically unbiased with vanishing variance: if $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$ and $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$, then $\hat{\theta}_n$ is consistent (this follows from Chebyshev's inequality). As the sample grows, a consistent estimator converges to the true parameter.

A single number combines the two concerns behind unbiasedness and efficiency — being centered on θ and being tightly concentrated.

DEFINITION 7.9 • Mean squared error

The **mean squared error** of an estimator $\hat{\theta}$ of θ is the average squared distance from the target,

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{Var}(\hat{\theta}) + B(\hat{\theta})^2,$$

the sum of its variance and its squared bias.

The decomposition follows by adding and subtracting $\mathbb{E}(\hat{\theta})$ inside the square; the cross term vanishes because $\mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})] = 0$. It shows that an unbiased estimator need not be best: a little bias can be worth accepting if it buys a large drop in variance.

EXAMPLE 7.7 • A biased estimator with smaller error

Compare the unbiased sample variance S^2 (divisor $n-1$) with the plug-in estimator $\hat{\sigma}_n^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2 = \frac{n-1}{n} S^2$ (divisor n). For a normal sample, $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$, and scaling gives

$$\mathrm{MSE}(S^2) = \frac{2\sigma^4}{n-1}, \qquad \mathrm{MSE}(\hat{\sigma}_n^2) = \underbrace{\frac{2(n-1)\sigma^4}{n^2}}_{\text{variance}} + \underbrace{\frac{\sigma^4}{n^2}}_{\text{bias}^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

At $n = 10$ these are $0.222\,\sigma^4$ and $0.190\,\sigma^4$: the *biased* estimator has the smaller mean squared error. Dividing by n tilts the estimate downward but shrinks its variance enough to win on total error.

REMARK 7.2 • The bias–variance trade-off in machine learning

This trade-off is the statistical heart of model fitting. A flexible model fit to minimize training error is nearly unbiased but high-variance — it overfits; deliberately biasing it through regularization, shrinkage, or a simpler hypothesis class raises bias a little while cutting variance, and can lower the expected prediction error overall. Choosing how much to regularize is, in these terms, choosing the point that minimizes mean squared error.

GUIDED PRACTICE 7.5

Let X_1, X_2, X_3, X_4 be a random sample with mean μ and variance σ^2, and let $T = (X_1 + 2X_2 + 3X_3 + 2X_4)/8$. Is T unbiased for μ, and is it more or less efficient than the sample mean $\overline{X}$? *Answer:* the weights sum to 8, so $\mathbb{E}(T) = (1+2+3+2)\mu/8 = \mu$ — unbiased. Its variance is $\mathrm{Var}(T) = (1^2+2^2+3^2+2^2)\sigma^2/8^2 = 18\sigma^2/64 = 0.281\sigma^2$, larger than $\mathrm{Var}(\overline{X}) = \sigma^2/4 = 0.25\sigma^2$. The equally weighted mean is more efficient; spreading weight unevenly inflates the variance.

7.2.3 Properties of the common estimators

We close by confirming that the three estimators of Section 7.1 have these virtues.

Sample mean. From the sample-mean result, $\mathbb{E}[\overline{X}] = \mu$, so $\overline{X}$ is unbiased. And since $\mathrm{Var}(\overline{X}) = \sigma^2/n \to 0$ as $n \to \infty$, it is also consistent for μ.

Sample proportion. Likewise $\mathbb{E}(\hat{p}) = p$ and $\mathrm{Var}(\hat{p}) = p(1-p)/n \to 0$, so $\hat{p}$ is unbiased and consistent for p.

Sample variance. The divisor $n-1$ is exactly what makes S^2 unbiased. To see it, write $(n-1)\mathbb{E}[S^2] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$ and use the decomposition

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2,$$

whose cross term vanishes because $\sum_i (X_i - \overline{X}) = 0$. Taking expectations,

$$(n-1)\mathbb{E}[S^2] = \sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2] - n\,\mathbb{E}[(\overline{X}-\mu)^2] = n\sigma^2 - n\cdot\frac{\sigma^2}{n} = (n-1)\sigma^2,$$

so $\mathbb{E}[S^2] = \sigma^2$. Its variance also tends to 0 as $n \to \infty$, so S^2 is unbiased and consistent for σ^2. (Note that dividing instead by n would make the estimator biased downward, because the data fall closer to their own mean $\overline{X}$ than to μ.)

EXERCISES (§7.2)

1. A random sample X_1, X_2, X_3 is drawn from a population with unknown mean μ and known variance σ^2. Consider the following estimators of μ:

$$\hat{\mu}_1 = \frac{X_1 + X_2 + X_3}{3}, \quad \hat{\mu}_2 = \frac{X_1 + 2X_2 + X_3}{4}, \quad \hat{\mu}_3 = X_3.$$

   Answer the following questions.

   (a) Show that $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ are each unbiased estimators of μ.

   (b) Find the most efficient estimator among $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$. (Hint: $\mathrm{Var}(\hat{\mu}_1) = \sigma^2/3$.)

2. Let X_1, X_2, X_3, X_4, X_5 be a random sample from a population with mean μ and variance σ^2. Show that the estimator

$$\hat{\mu} = \frac{2X_1 + X_2 + 2X_3 + X_4 + X_5}{7}$$

is an unbiased estimator of μ.

3. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed Bernoulli random variables with parameter p, i.e. $X_i \sim B(1, p)$. Answer the following.

(a) Consider the estimators $\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{p}_2 = \frac{1}{2}(X_1 + X_n)$. Determine whether each estimator is unbiased.

(b) Determine which of $\hat{p}_1$ or $\hat{p}_2$ is more efficient (for $n > 2$).

(c) Show that $\hat{p}_1$ is a consistent estimator of p.

(d) Derive a 95% confidence interval for p, assuming the sample size n is sufficiently large.

4. Suppose we obtain a random sample $X_1, X_2, \ldots, X_n$ from a population with probability density function

$$f(x) = \begin{cases} \frac{2}{\theta}x, & 0 < x < \sqrt{\theta}, \quad (\theta > 0), \\ 0, & \text{otherwise}, \end{cases}$$

where θ is an unknown parameter. $\left(\text{Hint: } E(X_1^2) = \frac{\theta}{2} \text{ and } \text{Var}(X_1^2) = \frac{\theta^2}{12}.\right)$

(a) Find the constant k such that the estimator $\hat{\theta} = \frac{k}{n}\sum_{i=1}^{n} X_i^2$ is an unbiased estimator of θ.

(b) Using the value of k from part (a), show that $\hat{\theta}$ is a consistent estimator of θ.

5. Suppose we obtain a random sample $X_1, X_2, \ldots, X_n$ from the population with probability density function

$$f(x) = \begin{cases} \frac{3x^2}{\theta^3}, & 0 < x < \theta \quad (\theta > 0), \\ 0, & \text{otherwise}. \end{cases}$$

For the estimator $\hat{\theta} = \frac{4}{3}\overline{X}_n = \frac{4}{3n}\sum_{i=1}^{n} X_i$, answer the following.

(a) Show that the estimator $\hat{\theta}$ is an unbiased estimator of θ.

(b) Show that the estimator $\hat{\theta}$ is a consistent estimator of θ.

(c) From a random sample of size $n = 60$, the observed sample mean is $\overline{x} = 3$. Construct an approximate 95% confidence interval for θ. (Hint: consider the sampling distribution of $\hat{\theta}$ using the Central Limit Theorem.)

CHAPTER 8

Inference for Means

Numerical data are most often summarized by a mean, so the natural target for inference is a population mean μ. This chapter develops confidence intervals and hypothesis tests for μ, and the central new idea is a practical one: the population standard deviation σ is almost never known, and estimating it with the sample standard deviation s replaces the normal model with Student's t**-distribution**. Section 8.1 builds confidence intervals for a single mean, Section 8.2 the corresponding tests, and Section 8.3 studies a test's **power**. Later sections extend the same machinery to paired samples, to the difference of two means, and to the comparison of several means at once. Throughout, the sampling distribution of $\overline{X}$ and the chi-squared distribution of Chapter 7 do the work.

8.1 Confidence interval for a population mean

8.1.1 The z-interval when σ is known

Recall from Chapter 7 that for a random sample from a normal population, $\overline{X} \sim N(\mu, \sigma^2/n)$, so the standardized mean

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

The middle $100(1-\alpha)\%$ of a standard normal lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$, where $z_{\alpha/2}$ is the value with upper-tail area $\alpha/2$ (Figure 8.1):

$$\mathbb{P}(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha.$$

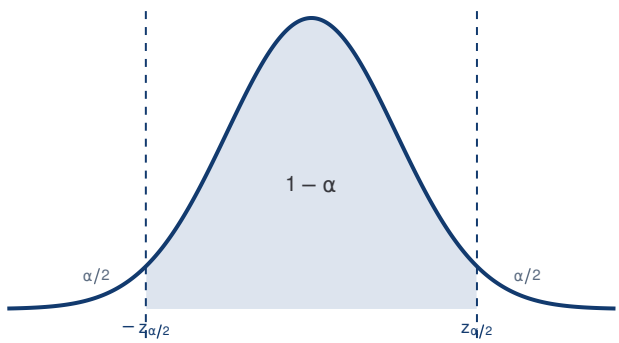


Figure 8.1. The middle $1-\alpha$ of the standard normal sits between $-z_{\alpha/2}$ and $z_{\alpha/2}$, leaving area $\alpha/2$ in each tail.

Substituting $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ and rearranging the inequality to isolate μ turns this statement about Z into one about μ:

$$1-\alpha = \mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;\le\; \mu \;\le\; \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

This is the same "point estimate ± critical value × SE" recipe as in Chapter 6, now with point estimate $\bar{x}$ and $SE = \sigma/\sqrt{n}$.

z-CONFIDENCE INTERVAL FOR A MEAN (KNOWN σ)

When σ is known and $\overline{X}$ is normal, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} \;\pm\; z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

EXAMPLE 8.1 • A known-σ interval for mean latency

An inference service measures the latency of $n = 49$ requests and finds a sample mean of $\bar{x} = 500.4$ ms. From long experience the population standard deviation is taken as known, $\sigma = 2.1$ ms. A 95% interval uses $z_{0.025} = 1.96$ and

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2.1}{\sqrt{49}} = 0.3, \qquad 500.4 \pm 1.96(0.3) \;\rightarrow\; (499.81,\ 500.99) \text{ ms.}$$

GUIDED PRACTICE 8.1

With σ known to be 9 ms, a service's latency is measured on $n = 36$ requests, giving $\bar{x} = 148$ ms. Construct a 95% confidence interval for the mean latency ($z_{0.025} = 1.96$).
Answer: $SE = 9/\sqrt{36} = 1.5$, so the interval is $148 \pm 1.96(1.5) = 148 \pm 2.94 = (145.06,\ 150.94)$ ms.

8.1.2 The t-distribution

In practice σ is rarely known, so we estimate it with the sample standard deviation s and use $SE \approx s/\sqrt{n}$. For large n this substitution barely matters, but for small n it injects real extra uncertainty, because s itself varies from sample to sample. To account for it we replace the normal model with the **t-distribution**.

The t-distribution has the same symmetric bell shape as $N(0,1)$ but *thicker tails*, reflecting the added uncertainty from estimating σ. It is centered at zero and has a single parameter, the **degrees of freedom** df; we write $T \sim t(df)$. The tails are heaviest when df is small, and as df grows the extra uncertainty fades and the distribution approaches $N(0,1)$ (Figure 8.2).

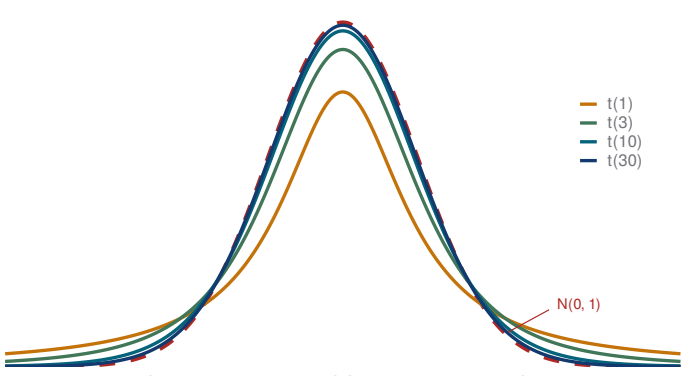


Figure 8.2. The t-distribution for several degrees of freedom, with the standard normal (dashed red) as its limit. Small df produces noticeably thicker tails and a lower peak; as df grows the curves converge to $N(0,1)$.

DEFINITION 8.1 • The t-distribution

If $Z \sim N(0,1)$ and $V \sim \chi^2(k)$ are independent, then

$$\frac{Z}{\sqrt{V/k}} \sim t(k),$$

the t-distribution with k degrees of freedom.

NOTE

This definition is exactly what the standardized sample mean requires. For an i.i.d. normal sample, Chapter 7 gives $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ *independently*, so

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{V/(n-1)}} \sim t(n-1).$$

Replacing the known σ by the estimate S is precisely what turns the $N(0,1)$ into a $t(n-1)$.

EXAMPLE 8.2 • A t tail area

What proportion of the $t(18)$ distribution falls below -2.10? Because the t is symmetric and tabulated, this is a single software call,

$$\mathbb{P}(T \le -2.10) = 0.025 \qquad (T \sim t(18)).$$

```
> pt(-2.10, df = 18)
[1] 0.0250452
```

EXAMPLE 8.3 • Reading a two-sided t critical value

For a 95% interval on $n = 15$ observations we need the upper-0.025 quantile of $t(14)$. It is slightly larger than the normal value $z_{0.025} = 1.96$, reflecting the thicker tails:

```
> qt(0.975, df = 14)
[1] 2.144787
```

REMARK 8.1 • The t critical value always exceeds the z value

Because the t has heavier tails than $N(0,1)$, the critical value $t_{\alpha/2}(df)$ is larger than $z_{\alpha/2}$, and the gap shrinks as df grows. At 95% confidence:

df	5	10	20	30	60	120	∞
$t_{0.025}(df)$	2.571	2.228	2.086	2.042	2.000	1.980	1.960

The practical reading: with small samples the extra width is real and worth carrying; past $df \approx 30$ the t- and z-intervals nearly coincide.

8.1.3 The t-interval when σ is unknown

Swapping σ for s and the normal critical value for a t critical value gives the interval used in almost all real applications.

t-CONFIDENCE INTERVAL FOR A MEAN (UNKNOWN σ)

From a sample of n independent, nearly normal observations, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} \pm t_{\alpha/2}(n-1)\,\frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}(n-1)$ is the upper-$\alpha/2$ quantile of the $t(n-1)$ distribution.

	z-interval	t-interval
Condition	σ known	σ unknown
SE	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
Critical value	$z_{\alpha/2}$	$t_{\alpha/2}(n-1)$

EXAMPLE 8.4 • Mean energy per inference

An accelerator's energy use is measured on $n = 19$ randomly chosen inference runs, giving $\bar{x} = 4.4$ J and $s = 2.3$ J. With σ unknown we use a t-interval. The standard error and degrees of freedom are

$$SE = \frac{s}{\sqrt{n}} = \frac{2.3}{\sqrt{19}} = 0.528, \qquad df = n - 1 = 18,$$

and the 95% critical value is $t_{0.025}(18) = 2.101$:

```
> qt(p = 0.025, df = 18, lower.tail = FALSE)
[1] 2.100922
```

Hence

$$4.4 \pm 2.101(0.528) \;\rightarrow\; (3.29,\ 5.51)\ \text{J}.$$

We are 95% confident the true mean energy per inference lies between 3.29 and 5.51 J.

A worked interval often starts not from summaries but from the raw numbers. The next example builds one from a vector of measurements, then varies the confidence level to expose a basic trade-off.

EXAMPLE 8.5 • A confidence interval from raw data

A trained classifier is evaluated on 10 disjoint, freshly drawn test batches; the batch accuracies are

`0.905  0.882  0.931  0.918  0.864  0.899  0.927  0.873  0.911  0.890`

We first reduce the sample to the two quantities the interval needs — the mean and the standard deviation,

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0.900, \qquad s = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2} = 0.0226,$$

so that $SE = s/\sqrt{n} = 0.0226/\sqrt{10} = 0.00716$ on $df = 9$. With $t_{0.025}(9) = 2.262$ the 95% interval is

$$0.900 \pm 2.262(0.00716) \;\rightarrow\; (0.884,\ 0.916).$$

In R the same interval comes straight from `t.test`:

```
> acc <- c(0.905,0.882,0.931,0.918,0.864,0.899,0.927,0.873,0.911,0.890)
> t.test(acc)$conf.int
[1] 0.8838 0.9162
attr(,"conf.level")
[1] 0.95
```

! CAUTION

This interval is valid because the ten batch accuracies are *independent* — a fixed model scored on disjoint test data. If the ten numbers were instead the fold accuracies of a single k-fold cross-validation, they would *not* be independent: the folds share most of their training data, so an ordinary t-interval would understate the uncertainty. Independent test batches (as here), repeated or nested cross-validation, or a bootstrap are needed for valid interval estimates from cross-validated scores.

The confidence level is a dial we choose, and it is not free: demanding more confidence forces a wider interval. Holding the same data fixed and recomputing at 90%, 95%, and 99% changes only the critical value, and the interval grows with it (Figure 8.3):

$$90\%\text{: } (0.887,\ 0.913), \qquad 95\%\text{: } (0.884,\ 0.916), \qquad 99\%\text{: } (0.877,\ 0.923).$$

REMARK 8.2 • What "95% confident" means

The confidence level describes the *procedure*, not a single interval. If the whole experiment were repeated many times, about 95% of the intervals built this way would contain the true μ — the capture picture of Chapter 6 (Figure 6.3). For the one interval we actually computed,

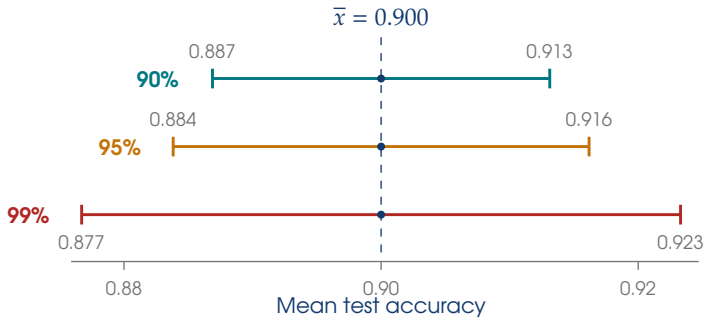


Figure 8.3. The same data give a wider interval at higher confidence. The point estimate $\bar{x} = 0.900$ is fixed; raising the confidence level from 90% to 99% enlarges the critical value, and the interval stretches outward around it.

μ is either inside it or not; it is the long-run success rate of the recipe that is 95%. The interval also says nothing about where individual batch accuracies fall — only about their mean.

GUIDED PRACTICE 8.2

Using $\bar{x} = 0.900$, $s = 0.0226$, and $n = 10$ from Example 8.5, build a 99% confidence interval for the mean accuracy, given $t_{0.005}(9) = 3.250$. Is it wider or narrower than the 95% interval, and why?
Answer: $SE = 0.00716$, so the margin is $3.250(0.00716) = 0.0233$ and the interval is $0.900 \pm 0.023 = (0.877, 0.923)$. It is wider than the 95% interval because a larger critical value is needed to capture μ with greater long-run frequency.

A confidence interval also answers a planning question asked *before* any data are collected: how large a sample is needed to pin the mean down to within a stated tolerance? The half-width of the z-interval,

$$m = z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},$$

is the **margin of error**. Fixing a target m and solving for n turns the interval formula into a sample-size rule.

SAMPLE SIZE FOR A DESIRED MARGIN OF ERROR

To estimate a mean to within a margin of error m with $100(1-\alpha)\%$ confidence, using a planning value σ, take

$$n \geq \left(\frac{z_{\alpha/2}\,\sigma}{m}\right)^2,$$

rounding *up* to the next whole number.

EXAMPLE 8.6 • Sizing a latency study

Engineers want a 95% interval for mean latency with a margin of error no larger than $m = 0.5$ ms, and from past monitoring take $\sigma = 2.1$ ms. With $z_{0.025} = 1.96$,

$$n \geq \left(\frac{1.96 \times 2.1}{0.5}\right)^2 = (8.232)^2 = 67.8 \;\rightarrow\; n = 68.$$

Rounding up matters: at $n = 67$ the margin is 0.503 ms, just over target, while $n = 68$ brings it to 0.499 ms. Halving the margin to 0.25 ms quadruples the requirement to $n = 272$, because n grows with the *square* of $1/m$ — precision is expensive.

GUIDED PRACTICE 8.3

Repeat the calculation for a 90% interval with the same $\sigma = 2.1$ ms and margin $m = 0.5$ ms, using $z_{0.05} = 1.645$.
Answer: $n \geq (1.645 \times 2.1/0.5)^2 = (6.909)^2 = 47.7$, so $n = 48$ — fewer than the 68 needed at 95%, because a lower confidence level uses a smaller critical value.

EXERCISES (§8.1)

1. A factory wants to estimate the mean operating time of machines μ. A random sample of $n = 15$ machines was selected, and their operating times (in minutes) were measured as follows:

i	1	2	3	4	5	6	$\cdots$	14	15
x_i	304.71	293.02	296.60	301.85	294.92	299.64	$\cdots$	308.23	297.21

   The sample mean is $\bar{x} = 299.16$, and the sample standard deviation is $s = 4.44$. Assuming that all the necessary conditions are satisfied, construct a 90% confidence interval for the mean operating time μ.

2. The weight (g) of soap bars produced at a company follows a normal distribution. A random sample of 25 bars was selected, and the sample mean was $\bar{x} = 97$. Answer the following.
   (a) If the population standard deviation is $\sigma = 2$, find a 95% confidence interval for the population mean μ.
   (b) If the sample standard deviation is $s = 2.0616$, find a 90% confidence interval for the population mean μ.

3. During a physical fitness test, a random sample of $n = 10$ cadets was selected. The number of push-ups completed in 2 minutes was recorded; the sample mean was $\bar{x} = 58$ and the sample standard deviation was $s = 8$. Construct a 95% confidence interval for the mean number of push-ups completed by all cadets in 2 minutes. Assume that the normality condition is met.

4. Read each statement and choose the most appropriate word to complete the sentence.
   (a) As the sample size increases, the variance of the sample mean becomes (larger / smaller).
   (b) The t distribution is a (left-skewed / symmetric / right-skewed) distribution.
   (c) Compared to the standard normal (Z) distribution, the t distribution has (thicker / thinner) tails.
   (d) As the degrees of freedom increase, the tails of the t distribution become (thicker / thinner).

8.2 Hypothesis test for a population mean

A confidence interval answers "what is μ?" We now take up the complementary question: do the data provide strong enough evidence to support a specific *claim* about μ? Answering it is the job of a **hypothesis test**, which pits a skeptical default against a competing claim and asks whether the evidence is strong enough to abandon the default.

8.2.1 The hypothesis-testing framework

DEFINITION 8.2 • Null and alternative hypotheses

The **null hypothesis** H_0 is the skeptical, status-quo position — typically that nothing unusual is happening. The **alternative hypothesis** H_A is the claim we are gathering evidence for. For a mean, H_0 fixes μ at a specific **null value** μ_0, and we reject H_0 only when the data are hard to explain under it.

EXAMPLE 8.7 • Stating the hypotheses

A serving system's mean latency is documented as $\mu_0 = 200$ ms, and a team suspects a recent change pushed it *higher*. The skeptical default is that the mean is unchanged, and the concern points in a single direction, so

$$H_0 : \mu = 200 \qquad \text{versus} \qquad H_A : \mu > 200.$$

This is a **one-sided** test; we carry it out in Example 8.8 once the test statistic is in hand.

The form of H_A depends on the question (Figure 8.4). A **two-sided** alternative $H_A : \mu \neq \mu_0$ looks for a departure in either direction and uses both tails. A **one-sided** alternative looks in a single direction: upper-tailed ($H_A : \mu > \mu_0$) or lower-tailed ($H_A : \mu < \mu_0$), using the corresponding tail.

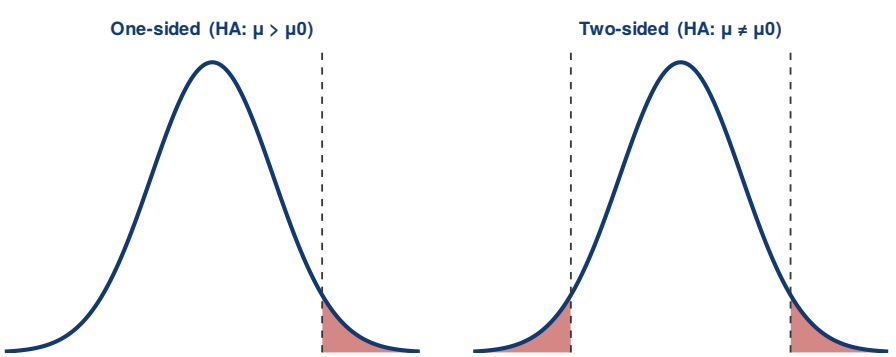


Figure 8.4. The alternative fixes where the evidence against H_0 lies. A one-sided alternative places the p-value in a single tail (left); a two-sided alternative splits it across both tails, doubling the area (right).

The reasoning mirrors a courtroom. A defendant is presumed innocent (H_0) and convicted (H_A) only on strong evidence. A failure to convict does not prove innocence — it means the evidence was insufficient. In the same way, failing to reject H_0 never proves H_0; it only means the data did not make a convincing case against it.

> **! CAUTION**
>
> "Fail to reject H_0" is not the same as "accept H_0." A test can only find evidence *against* the null; the absence of such evidence is not positive proof that the null is true.

8.2.2 Decision errors and the significance level

A hypothesis test reaches a verdict, and like any verdict it can be wrong in two ways.

> **DEFINITION 8.3 • Type 1 and Type 2 errors**
>
> A **Type 1 error** is rejecting H_0 when it is in fact true. A **Type 2 error** is failing to reject H_0 when H_A is in fact true.

	Decision	
Truth	fail to reject H_0	reject H_0
H_0 true	correct	Type 1 error
H_A true	Type 2 error	correct

The two error rates are conditional probabilities, written α and β:

$$\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true}), \qquad \beta = \mathbb{P}(\text{fail to reject } H_0 \mid H_A \text{ true}).$$

> **DEFINITION 8.4 • Significance level**
>
> The **significance level** α is the largest Type 1 error rate we are willing to tolerate, fixed in advance and conventionally $\alpha = 0.05$. A test rejects H_0 exactly when the evidence against it would arise with probability at most α if H_0 were true.

> **! CAUTION**
>
> The two error rates trade off: tightening α to reduce Type 1 errors enlarges β, and loosening it does the reverse. The only way to shrink both at once is to collect more data, which narrows both sampling distributions.

Figure 8.5 shows both error types at once for a one-sided test. A cutoff c splits the sample mean into a fail-to-reject region (below c) and a reject region (above c). The null's mass in the reject region is

the Type 1 rate α; the alternative's mass in the fail-to-reject region is the Type 2 rate β. Sliding c to shrink one region enlarges the other.

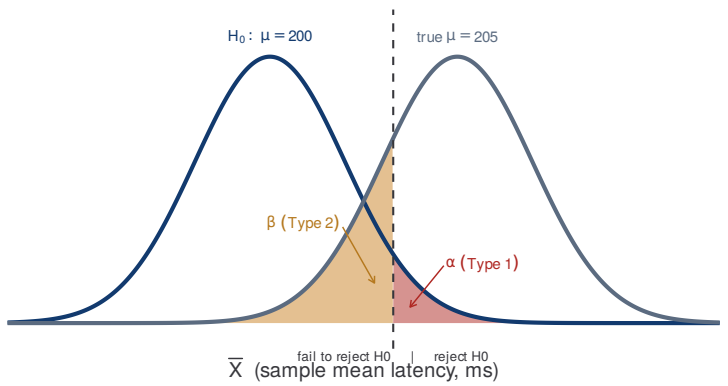


Figure 8.5. Two competing distributions of the sample mean $\overline{X}$: under H_0 ($\mu = 200$) and under one possible truth ($\mu = 205$). The cutoff (dashed) splits the decision. The Type 1 region (α) is the null's tail in the reject zone; the Type 2 region (β) is the alternative's mass in the fail-to-reject zone.

The complement of the Type 2 rate, $1 - \beta$, is the probability of *correctly* rejecting H_0 when H_A holds — the **power** of the test. Power rises with the sample size and with the size of the effect; Section 8.3 computes it in full.

8.2.3 The z-test (known σ)

To judge the evidence we ask: if H_0 were true, how surprising would the data be? When σ is known, the sample mean is normal under H_0, $\overline{X} \overset{H_0}{\sim} N(\mu_0, \sigma^2/n)$, so the standardized statistic is standard normal.

ONE-SAMPLE z-TEST FOR A MEAN

For $H_0 : \mu = \mu_0$ with known σ, the test statistic and its null distribution are

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1).$$

Alternative	Rejection region	p-value
$H_A : \mu \neq \mu_0$	$\lvert z\rvert > z_{\alpha/2}$	$2\,\mathbb{P}(Z > \lvert z\rvert)$
$H_A : \mu > \mu_0$	$z > z_\alpha$	$\mathbb{P}(Z > z)$
$H_A : \mu < \mu_0$	$z < -z_\alpha$	$\mathbb{P}(Z < z)$

The decision can be framed in two equivalent ways. The first measures how extreme the observed statistic is.

DEFINITION 8.5 • p-value

The **p-value** is the probability, computed *assuming* H_0 *is true*, of obtaining a test statistic at least as extreme as the one observed, in the direction favored by H_A. A small p-value means the data would be unlikely under H_0, and so counts as evidence against it; the decision rule is to reject H_0 when the p-value is at most α.

EXAMPLE 8.8 • Has mean latency risen?

We carry out the latency test of Example 8.7: $H_0 : \mu = 200$ against $H_A : \mu > 200$, with known $\sigma = 12$ ms. On $n = 36$ requests the sample mean is $\bar{x} = 204$ ms, so

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{204 - 200}{12/\sqrt{36}} = \frac{4}{2} = 2.00.$$

Because H_A is upper-tailed, only large values count as evidence, so the p-value is the upper-tail area (Figure 8.6):

$$\text{p-value} = \mathbb{P}(Z \geq 2.00) = 0.023.$$

Since $0.023 < 0.05$, we reject H_0: the data give strong evidence that mean latency now exceeds 200 ms.

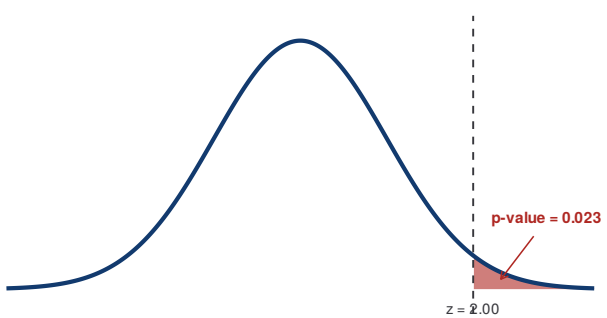


Figure 8.6. The p-value for the latency test. Under H_0 the statistic is standard normal; the shaded upper tail beyond the observed $z = 2.00$ has area 0.023.

The second framing fixes a cutoff in advance.

DEFINITION 8.6 • Rejection region

The **rejection region** is the set of test-statistic values for which H_0 is rejected at level α, and its boundary is the **critical value**. For the upper-tailed test the critical value is z_α, chosen so that $\mathbb{P}(Z > z_\alpha) = \alpha$.

For the latency test the rejection region is $z > z_{0.05} = 1.645$, and the observed $z = 2.00$ falls inside it — the same verdict the p-value gave.

NOTE

The rejection-region and p-value approaches always reach the same conclusion: the observed statistic lands in the rejection region exactly when the p-value is at most α.

8.2.4 The t-test (unknown σ)

When σ is unknown — the usual case — we estimate it with s, and the null distribution becomes a t.

ONE-SAMPLE t-TEST FOR A MEAN

For $H_0 : \mu = \mu_0$ with unknown σ,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t(n-1).$$

Alternative	Rejection region	p-value
$H_A : \mu \neq \mu_0$	$\lvert t\rvert > t_{\alpha/2}(n-1)$	$2\,\mathbb{P}(T > \lvert t\rvert)$
$H_A : \mu > \mu_0$	$t > t_{\alpha}(n-1)$	$\mathbb{P}(T > t)$
$H_A : \mu < \mu_0$	$t < -t_{\alpha}(n-1)$	$\mathbb{P}(T < t)$

EXAMPLE 8.9 • Did mean response length change?

In an earlier release a model produced a mean of $\mu_0 = 93.29$ tokens per response. A team asks whether the mean has *changed* in the new release, with σ unknown. On $n = 100$ responses they observe $\overline{x} = 97.32$ and $s = 16.98$, so

$$t = \frac{97.32 - 93.29}{16.98/\sqrt{100}} = \frac{4.03}{1.698} = 2.37, \qquad df = 99.$$

Because $H_A : \mu \neq \mu_0$ is two-sided, both tails count:

$$\text{p-value} = 2\,\mathbb{P}(T \geq 2.37) = 2(0.0099) = 0.020.$$

Since $0.020 < 0.05$, we reject H_0: the mean response length differs from the earlier release. Equivalently, the rejection region is $\lvert t\rvert > t_{0.025}(99) = 1.984$, which the observed $\lvert t\rvert = 2.37$ exceeds. The same test in R:

```
> t.test(tokens, alternative = "two.sided", mu = 93.29)

 One Sample t-test

data: tokens
t = 2.3701, df = 99, p-value = 0.019721
alternative hypothesis: true mean is not equal to 93.29
95 percent confidence interval:
 93.93546 100.70783
sample estimates:
mean of x
 97.32167
```

The token example is two-sided. When the question carries a direction — has fine-tuning *reduced* latency? — the alternative is one-sided and the whole rejection region sits in a single tail.

EXAMPLE 8.10 • Did fine-tuning lower mean latency?

A model previously served requests at a mean latency of $\mu_0 = 250$ ms. After fine-tuning, a team measures $n = 16$ requests and finds $\bar{x} = 244$ ms with $s = 10$ ms. Because they ask only whether latency has *fallen*, the test is $H_0 : \mu = 250$ versus $H_A : \mu < 250$. The statistic is

$$t = \frac{244 - 250}{10/\sqrt{16}} = \frac{-6}{2.5} = -2.40, \qquad df = 15.$$

Only the lower tail counts, so

$$\text{p-value} = \mathbb{P}(T \le -2.40) = 0.015.$$

Since $0.015 < 0.05$ we reject H_0; equivalently the observed $t = -2.40$ falls below the one-sided critical value $-t_{0.05}(15) = -1.753$. The evidence supports a real drop in mean latency. Had the alternative been two-sided the p-value would have doubled to 0.030 — still significant here, but a one-sided test spends all of α on the direction of interest, so its critical value is less extreme and a true one-directional effect is detected with less data (Figure 8.7).

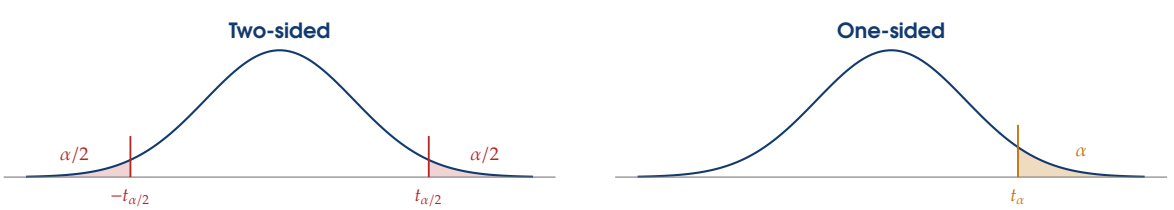


Figure 8.7. The same significance level α, split two ways. A two-sided test places $\alpha/2$ in each tail, so its critical value $t_{\alpha/2}$ is farther out; a one-sided test puts all of α in the tail of interest, pulling the critical value t_α closer to the center. Concentrating the whole error budget in one direction is what makes a correctly chosen one-sided test more powerful against an effect in that direction.

NOTE

As $n \to \infty$, $t(n-1) \to N(0,1)$ and the t-test converges to the z-test, but for any finite n the t critical value $t_{\alpha/2}(n-1)$ is larger than $z_{\alpha/2}$, demanding slightly stronger evidence to compensate for the estimation of σ.

NOTE

A two-sided test and a confidence interval are two views of the same evidence. A level-α two-sided test of $H_0 : \mu = \mu_0$ rejects *exactly when* μ_0 falls outside the $100(1-\alpha)\%$ confidence interval. In Example 8.9 the 95% interval is (93.94, 100.71), which does not contain the null value 93.29 — the same rejection the p-value gave at $\alpha = 0.05$.

GUIDED PRACTICE 8.4

A monitoring script records the CPU temperature (°C) of a server at five random times: $61, 63, 60, 64, 62$. Test $H_0 : \mu = 60$ against $H_A : \mu \neq 60$ at $\alpha = 0.05$, given $t_{0.025}(4) = 2.776$.
Answer: $\bar{x} = 62$, $s = \sqrt{2.5} = 1.581$, and $SE = 1.581/\sqrt{5} = 0.707$, so $t = (62-60)/0.707 = 2.83$ on $df = 4$. Since $|t| = 2.83 > 2.776$ (p-value ≈ 0.047), we reject H_0: the mean temperature differs from 60°C, though only narrowly.

REMARK 8.3 • Statistical vs. practical significance, and effect size

A small p-value says an effect is unlikely to be exactly zero, not that it is large: with enough data even a trivial difference becomes "significant." A complementary, sample-size-free summary is the standardized **effect size**, for one mean

$$d = \frac{\bar{x} - \mu_0}{s},$$

the shift measured in standard deviations. In Example 8.9 the change of about 4 tokens against $s = 16.98$ gives $d = 4.03/16.98 \approx 0.24$ — statistically clear at $n = 100$, yet only a quarter of a standard deviation, a modest practical effect. Report the estimate and its confidence interval alongside the p-value, never the p-value alone.

EXERCISES (§8.2)

1. A company produces steel components whose average breaking strength is known to be $\mu_0 = 655$ MPa. An outside laboratory claimed that the true average breaking strength is now below 655 MPa. To investigate this claim, $n = 9$ steel components are randomly selected from the production line and their breaking strengths $X_1, X_2, \ldots, X_9$ are measured. The company wishes to test this claim at significance level $\alpha = 0.05$. Assume that breaking strength follows a normal distribution.
   (a) State the null and alternative hypotheses.
   (b) Find the test statistic and its null distribution.

(c) The breaking strengths (MPa) of the 9 randomly selected steel components are given below. Find the observed value of the test statistic.

Component (i)	1	2	3	$\cdots$	6	7	8	9
Breaking strength (x_i, MPa)	621	676	634	$\cdots$	678	651	646	620

Note: $\sum_{i=1}^{9} x_i = 5832$, $\sum_{i=1}^{9} (x_i - \overline{x})^2 = 3528$

(d) Using the result in part (c), report the p-value and complete the hypothesis test. State the conclusion in the context of the data.

2. A hat manufacturer wants to check whether the average head size of new recruits is 24 inches. To investigate, 25 recruits were randomly selected and their head sizes $X_1, X_2, \ldots, X_{25}$ were measured. The manufacturer wishes to test $H_0 : \mu = 24$ vs. $H_A : \mu \neq 24$ at significance level $\alpha = 0.05$. Assume that head size follows a normal distribution with variance $\sigma^2 = 4$. The test statistic is $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, where $\overline{X}$ is the sample mean, μ_0 is the null value, and n is the sample size.

(a) Find the null distribution of the test statistic.

(b) Find the rejection region at significance level $\alpha = 0.05$.

(c) The sample mean of the 25 selected recruits is $\overline{x} = 25$ inches. Using the rejection region in part (b), complete the hypothesis test. State the conclusion in the context of the data.

3. A buyer wants to test whether the average diameter of nuts from a supplier is different from $\mu_0 = 12.7$ mm. A random sample of $n = 10$ nuts is selected and their diameters $X_1, \ldots, X_{10}$ are measured. Test at significance level $\alpha = 0.05$. Assume the nut diameters follow a normal distribution.

(a) State the null and alternative hypotheses.

(b) Find the test statistic and its null distribution.

(c) The measured diameters of the 10 selected nuts are shown below. Find the observed value of the test statistic.

Nut (i)	1	2	3	$\cdots$	8	9	10
Diameter (x_i, mm)	12.68	12.79	12.85	$\cdots$	12.67	12.79	12.65

$\overline{x} = 12.744, \quad s = 0.0783.$

(d) Using the result in part (c), report the p-value and complete the hypothesis test. State the conclusion in the context of the data.

4. The average recovery time for an existing health supplement is $\mu_0 = 20$ days. A team claims that their new supplement gives a smaller average recovery time (less than 20 days). To test this, $n = 16$ people take the new supplement and their recovery times $X_1, \ldots, X_{16}$ are measured. Test at significance level $\alpha = 0.05$. Assume the recovery time follows a normal distribution with standard deviation $\sigma = 4$ days.

(a) State the null and alternative hypotheses.

(b) Find the test statistic and its null distribution.

(c) Find the rejection region at significance level $\alpha = 0.05$.

(d) The sample mean recovery time is $\overline{x} = 18.2$ days. Find the observed value of the test statistic.

(e) Using the results in part (d) and the rejection region, complete the hypothesis test. State the conclusion in the context of the data.

5. The average storage capacity of USB drives produced by Company A, denoted μ, used to be reported as 16 GB. Recently, some consumers have claimed that the average capacity is *less* than 16 GB. To verify this claim, a random sample of 9 USB drives was selected and their capacities (in GB) were measured. Conduct *a hypothesis test for a population mean* at significance level $\alpha = 0.05$. (Assume that the storage capacities follow a normal distribution.)

(a) State the null and alternative hypotheses. (Use a one-sided test.)

(b) Find the test statistic and its null distribution.

(c) The sample mean and sample standard deviation from the 9 USB drives are $\bar{x} = 15.84$ and $s = 0.24$. Compute the observed test statistic.

(d) Compute the p-value and complete the hypothesis test. State the conclusion in the context of the data. (Use $\mathbb{P}(T < -2) \approx 0.0403$ for $T \sim t(8)$.)

8.3 Power calculations for a population mean

A test can fail to detect a real effect. The **power** of a test is its probability of correctly rejecting H_0 when a specific alternative is true:

$$\text{Power} = \mathbb{P}(\text{reject } H_0 \mid H_A \text{ true}) = 1 - \beta.$$

Power rises with a larger sample size, a larger true effect, and a larger α; the catch is that lowering α to avoid Type 1 errors directly lowers power. Computing power for a mean is a concrete three-step exercise: find the rejection region in the units of $\overline{X}$, suppose a specific true mean, and compute the chance that $\overline{X}$ lands in the rejection region.

EXAMPLE 8.11 • Power of a delivery-time test

A courier's mean delivery time is $\mu_0 = 30$ minutes with known $\sigma = 5$. After a new training program the manager tests whether the mean has *dropped*, $H_0 : \mu = 30$ versus $H_A : \mu < 30$, at $\alpha = 0.05$ with $n = 25$, so $SE = \sigma/\sqrt{n} = 1$.
Step 1 — rejection region in $\overline{X}$. Small means are evidence against H_0, so we reject when $\bar{x} < c$. Standardizing, $\mathbb{P}(\overline{X} < c \mid \mu_0) = 0.05$ requires $(c - 30)/1 = -z_{0.05} = -1.645$, hence

$$c = 30 - 1.645(1) = 28.355.$$

Step 2 — suppose a true mean. If the program truly lowered the mean to $\mu = 27$, then $\overline{X} \overset{H_A}{\sim} N(27, 1^2)$.
Step 3 — probability of landing in the rejection region.

$$\text{Power} = \mathbb{P}(\overline{X} < 28.355 \mid \mu = 27) = \mathbb{P}\left(Z < \frac{28.355 - 27}{1}\right) = \mathbb{P}(Z < 1.355) \approx 0.91.$$

The test has about a 91% chance of detecting a true mean as low as 27 minutes (Figure 8.8).

! CAUTION

Power is always stated *relative to a specific alternative*. Detecting a large effect is easy and the power is high; detecting an effect just barely different from μ_0 is hard and the power is low. Power is a *prospective* planning quantity: it describes how reliably a study of a given size would detect a prespecified effect *before* any data are collected. After a study fails to reject

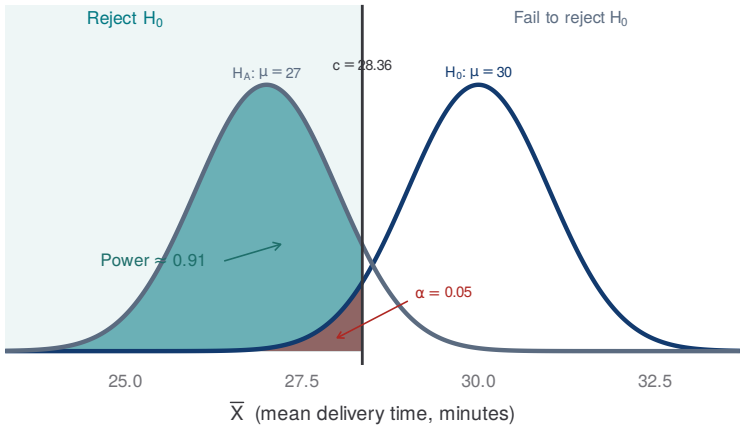


Figure 8.8. Power for the one-sided delivery-time test. The vertical cutoff at $c = 28.355$ defines the rejection region $\bar{x} < c$. The power is the shaded area of the alternative curve ($\mu = 27$) to the left of the cutoff; the small shaded area of the null curve there is α.

H_0, recomputing power from the observed estimate cannot tell us whether the effect is truly absent or merely too small for the sample to catch. Instead, report the confidence interval, which shows the range of effect sizes compatible with the data.

Power also answers a planning question asked *before* any data are collected: how large must n be to reach a target power against a stated effect? Inverting the power calculation gives a formula.

SAMPLE SIZE FOR A ONE-SIDED ONE-SAMPLE TEST

For $H_0 : \mu = \mu_0$ against a one-sided alternative with known σ, detecting a true mean μ_1 (effect $\Delta = |\mu_0 - \mu_1|$) with power $1 - \beta$ at level α requires

$$n \geq \left(\frac{(z_\alpha + z_\beta)\,\sigma}{\Delta} \right)^2 .$$

EXAMPLE 8.12 • Sizing the delivery-time study

Return to the courier of Example 8.11, with $\sigma = 5$, $\alpha = 0.05$, and the lower-tailed alternative. Suppose we want 90% power to detect a drop of $\Delta = 2$ minutes (from $\mu_0 = 30$ to $\mu_1 = 28$). With $z_{0.05} = 1.645$ and $z_{0.10} = 1.282$,

$$n \geq \left(\frac{(1.645 + 1.282)\,5}{2} \right)^2 = (7.32)^2 = 53.5,$$

so $n = 54$ runs suffice; a direct check gives power 0.90 at $n = 54$ and just under it at $n = 53$. Halving the effect we wish to detect would *quadruple* the required sample size, since n scales

with $1/\Delta^2$.

GUIDED PRACTICE 8.5

Keeping $\sigma = 5$ and $\alpha = 0.05$, how does the required n change if we only need 80% power ($z_{0.20} = 0.842$) to detect the same $\Delta = 2$?
Answer: $n \geq \left((1.645 + 0.842)\,5/2\right)^2 = (6.22)^2 = 38.7$, so $n = 39$. Less power is easier to achieve, so the study can be smaller.

EXERCISES (§8.3)

1. For the delivery-time test, recompute the power if the true mean is $\mu = 28$ instead of 27, using $\mathbb{P}(Z < 0.355) \approx 0.639$. Comment on why it is lower.

8.4 Paired data

So far each example has involved a single sample. Frequently, though, two measurements are taken on the *same* unit and we care about the difference between them — a before-and-after reading, or two methods applied to the same item. Such data are **paired**, and treating the two columns as independent would throw away the very structure that makes the comparison sharp.

8.4.1 Paired observations

DEFINITION 8.7 • Paired data

Two sets of observations are **paired** if each observation in one set corresponds to exactly one observation in the other.

Suppose two versions of a model, v1 and v2, are each run on the *same* $n = 68$ prompts, and for every prompt we record the latency difference d = (v1 latency) − (v2 latency). The two measurements on a prompt are not independent — a hard prompt is slow for both versions — so we collapse each pair to its single difference d_i and analyze those. Pairing cancels the prompt-to-prompt variation and isolates the version effect (Figure 8.9).

8.4.2 The paired t-test

Once we work with the differences, a paired test is nothing more than the one-sample t-test of Section 8.2 applied to $d_1, \ldots, d_{n_{\text{diff}}}$.

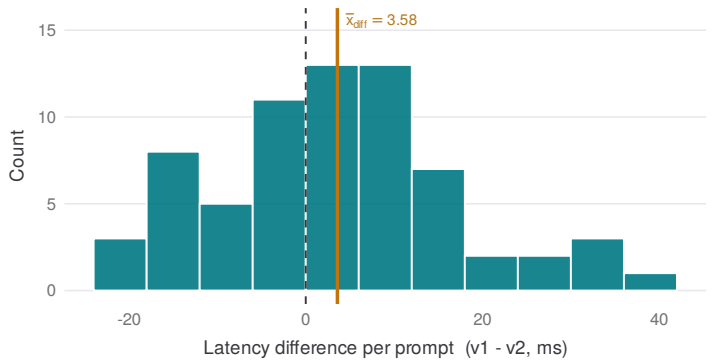


Figure 8.9. Per-prompt latency differences for two model versions run on the same prompts. The analysis reduces to the one sample of differences, whose mean $\bar{x}_{\text{diff}} = 3.58$ ms sits just above zero.

PAIRED t-TEST

For a test of the mean difference, $H_0 : \mu_{\text{diff}} = \delta_0$,

$$T = \frac{\overline{X}_{\text{diff}} - \delta_0}{S_{\text{diff}}/\sqrt{n_{\text{diff}}}} \overset{H_0}{\sim} t(n_{\text{diff}} - 1),$$

where $\overline{X}_{\text{diff}}$, S_{diff}, and n_{diff} are the mean, standard deviation, and number of the paired differences.

EXAMPLE 8.13 • Comparing two model versions

The $n_{\text{diff}} = 68$ latency differences have $\bar{x}_{\text{diff}} = 3.58$ ms and $s_{\text{diff}} = 13.42$ ms. We test $H_0 : \mu_{\text{diff}} = 0$ against $H_A : \mu_{\text{diff}} \neq 0$. The standard error and test statistic are

$$SE = \frac{s_{\text{diff}}}{\sqrt{n_{\text{diff}}}} = \frac{13.42}{\sqrt{68}} = 1.63, \qquad t = \frac{3.58 - 0}{1.63} = 2.20, \qquad df = 67.$$

Since the test is two-sided, p-value $= 2\,\mathbb{P}(T \geq 2.20) = 2(0.0156) = 0.031$. As $0.031 < 0.05$, we reject H_0: the two versions differ in mean latency. The same test in R:

```
> t.test(v1, v2, alternative = "two.sided", mu = 0, paired = TRUE)

 Paired t-test

data: v1 and v2
t = 2.2012, df = 67, p-value = 0.03117
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.3340641 6.8324064
sample estimates:
mean difference
 3.583235
```

The same differences also yield an interval. A confidence interval for μ_{diff} is just the one-sample t-interval applied to the d_i, and it carries the test's conclusion with it.

EXAMPLE 8.14 • Interval for the mean version difference

For the $n_{\text{diff}} = 68$ latency differences of Example 8.13, $\bar{x}_{\text{diff}} = 3.58$ ms, $s_{\text{diff}} = 13.42$ ms, and $SE = 1.63$ on $df = 67$. With $t_{0.025}(67) = 1.996$,

$$3.58 \pm 1.996(1.63) \;\rightarrow\; (0.33,\ 6.83) \text{ ms}.$$

The interval excludes 0 — matching the rejection in Example 8.13 — and is exactly the `95 percent confidence interval` printed by `t.test` there. Because it lies entirely above 0, v1 is the slower version, by somewhere between 0.3 and 6.8 ms on average.

REMARK 8.4 • Paired or independent?

The design, not the data, decides the test. Ask whether each value in one group is tied to a *specific* value in the other. Two model versions run on the *same* prompts are paired; two *different* sets of servers are independent. Treating paired data as independent is the costly error — it ignores the prompt-to-prompt variation that pairing removes, inflating the standard error and hiding effects that the paired analysis would reveal.

GUIDED PRACTICE 8.6

An optimization is applied to a service, and the latency (ms) of the same five requests is recorded before and after:

Before	120	115	130	125	110
After	118	110	126	119	107

Test whether the optimization *lowered* mean latency, $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} > 0$ for d = before − after, at $\alpha = 0.05$ with $t_{0.05}(4) = 2.13$.

Answer: the differences are $2, 5, 4, 6, 3$ with $\bar{x}_{\text{diff}} = 4$ and $s_{\text{diff}} = 1.58$, so $SE = 1.58/\sqrt{5} = 0.71$ and $t = 4/0.71 = 5.66$ on $df = 4$. Since $5.66 > 2.13$ (p-value ≈ 0.002), reject H_0: the optimization reduced mean latency.

EXERCISES (§8.4)

1. A trainer at Gym W claims that an exercise program increases muscle mass. For each of $n = 10$ participants, muscle mass (in kg) is measured before and after completing the program, and the difference is recorded. Let μ_{before} and μ_{after} denote the population mean muscle mass before and after the program, and define $\mu_{\text{diff}} = \mu_{\text{before}} - \mu_{\text{after}}$. A paired t-test is conducted at significance level $\alpha = 0.05$.

   (a) State the null and alternative hypotheses. (Use a one-sided test.)

   (b) Find the null distribution of the test statistic.

   (c) The before and after muscle mass measurements (in kg) are shown below. For the paired differences, the sample mean and sample standard deviation are $\bar{x}_{\text{diff}} = -1.8000$ and $s_{\text{diff}} = 2.9740$. Compute the observed test statistic.

Participant	1	2	3	4	5	6	7	8	9	10
Before	28	20	22	30	35	32	27	30	31	33
After	30	20	21	35	41	32	25	34	36	32
Diff. (Before − After)	−2	0	1	−5	−6	0	2	−4	−5	1

   (d) Compute the p-value and complete the hypothesis test. State the conclusion in the context of the data.

2. A company developed an 8-week diet program and claims that it reduces men's body fat. To evaluate this claim, the company measured the body fat of 9 male participants before and after completing the program. Let μ_{before} and μ_{after} denote the population mean body fat before and after the program, and let $\mu_{\text{diff}} = \mu_{\text{before}} - \mu_{\text{after}}$. Conduct a paired t-test at significance level $\alpha = 0.05$ to test whether the program effectively reduces body fat on average.

   (a) State the null and alternative hypotheses. (Use a one-sided test.)

   (b) Find the test statistic and its null distribution.

   (c) Using the R code and output provided, write down the p-value and complete the hypothesis test. State the conclusion in the context of the data.

```
before <- c(14, 15, 15, 17, 16, 16, 17, 15, 16)
after  <- c(12, 14, 12, 15, 14, 17, 13, 13, 17)
t.test(before, after, alternative = "greater", paired = TRUE)

  Paired t-test

data:  before and after
t = 2.8, df = [     ], p-value = [     ]
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 0.5224733       Inf
sample estimates:
mean difference
       1.555556
```

3. A newly developed protein supplement is tested to see whether it increases grip strength (kg). For $n = 12$ randomly selected people, grip strength is measured before, X_{Ai}, and after, X_{Bi} ($i = 1, \ldots, 12$), taking the supplement. Using a paired test at significance level $\alpha = 0.05$, we test whether the average grip strength is greater after than before. Assume the before, X_A, and after, X_B, grip strengths each follow a normal distribution with means μ_A and μ_B, and define $\mu_{\text{diff}} = \mu_A - \mu_B$. The hypotheses are $H_0 : \mu_{\text{diff}} = 0$ and $H_A : \mu_{\text{diff}} < 0$.

(a) Find the test statistic and its null distribution.

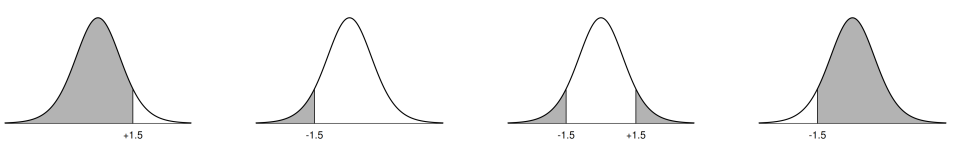
(b) Find the rejection region at significance level $\alpha = 0.05$. (Use $t_{0.05}(11) = 1.7959$.)

(c) The partial R output for this problem is shown below. Using the partial R output and the rejection region in part (b), complete the hypothesis test. State the conclusion in the context of the data.

```
  Paired t-test

data:  before and after
t = -1.5, df = [     ], p-value = [     ]
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
      -Inf 0.5917717
sample estimates:
mean difference
             -3
```

(d) Among the figures below, choose the one whose shaded (gray) region best corresponds to the p-value in part (c).



① ② ③ ④

8.5 Difference of two means: unequal variances

When the two groups are *independent* — different units in each, with no natural pairing — we compare them through the difference of sample means $\overline{X}_1 - \overline{X}_2$. Because the samples are independent, the variances add: with $\text{Var}(\overline{X}_1) = \sigma_1^2/n_1$ and $\text{Var}(\overline{X}_2) = \sigma_2^2/n_2$,

$$\text{Var}(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \qquad SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

estimating each σ_i^2 by S_i^2. Making no assumption that the two variances are equal gives the most broadly applicable test.

TWO-SAMPLE t-TEST (UNEQUAL VARIANCES)

For independent samples and $H_0 : \mu_1 - \mu_2 = \delta_0$,

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{H_0}{\sim} t(\psi),$$

with the Satterthwaite degrees of freedom

$$\psi = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

EXAMPLE 8.15 • Does a new caching strategy help?

Nine servers receive a new caching strategy and nine independent servers act as controls; each server's weekly change in throughput (requests per second) is recorded (Figure 8.10).

Group	n	$\bar{x}$	s
New caching	9	3.50	5.17
Control	9	−4.33	2.76

To test whether caching improves throughput we use a one-sided alternative, $H_0 : \mu_{\text{new}} - \mu_{\text{ctrl}} = 0$ versus $H_A : \mu_{\text{new}} - \mu_{\text{ctrl}} > 0$. The standard error and statistic are

$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95, \qquad t = \frac{3.50 - (-4.33)}{1.95} = 4.01,$$

and the Satterthwaite degrees of freedom work out to $\psi = 12.22$. The upper-tail p-value is $\mathbb{P}(T \geq 4.01) = 0.0008$, far below 0.05, so we reject H_0: the data give convincing evidence that the new caching strategy raises throughput. The Welch test in R:

```
> t.test(new, control, alternative = "greater", mu = 0)

 Welch Two Sample t-test

data: new and control
t = 4.0073, df = 12.225, p-value = 0.0008388
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.35469 Inf
sample estimates:
mean of x mean of y
 3.500000 -4.333333
```

The same Welch machinery yields an interval for the difference. Dropping the directional alternative and using a two-sided critical value turns the test into a 95% confidence interval for $\mu_{\text{new}} - \mu_{\text{ctrl}}$.

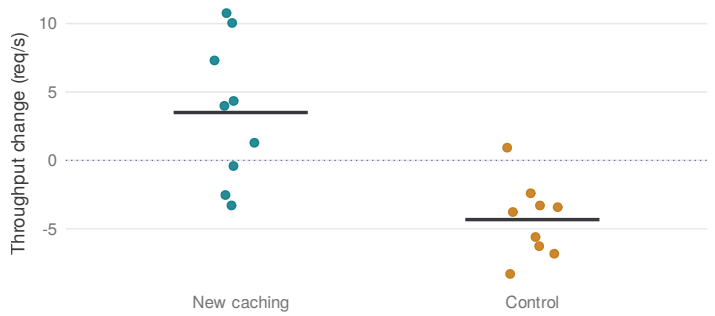


Figure 8.10. Weekly throughput change for the two independent groups, with each group mean marked. The new-caching group is centered well above the control, and the two groups show clearly different spreads.

EXAMPLE 8.16 • Interval for the throughput gain

From the caching data, $\bar{x}_{\text{new}} - \bar{x}_{\text{ctrl}} = 3.50 - (-4.33) = 7.83$ with $SE = 1.95$ on $\psi = 12.22$ Satterthwaite degrees of freedom. The two-sided critical value is $t_{0.025}(12.22) = 2.17$, so

$$7.83 \pm 2.17(1.95) \;\rightarrow\; (3.58,\ 12.08) \text{ rps.}$$

The interval lies entirely above zero, agreeing with the one-sided test's rejection of H_0, and it quantifies the gain: somewhere between about 3.6 and 12.1 additional requests per second. In R,

```
> t.test(new, control)$conf.int
[1]   3.58198 12.07802
attr(,"conf.level")
[1] 0.95
```

REMARK 8.5 • Welch or pooled?

The Welch interval makes no assumption that the two groups share a variance — and the sample SDs (5.17 versus 2.76) suggest they do not. When the spreads are comparable the pooled, equal-variance procedure of the next section is slightly more powerful, but Welch's test is the safer default: it keeps its error rate even under unequal variances, and costs little when the variances happen to match.

EXERCISES (§8.5)

1. The driving distance per full charge (in km) of an electric vehicle is known to follow a normal distribution: with a new battery (A), mean μ_A and variance $\sigma_A^2 = 18$; with the existing battery (B), mean μ_B and variance $\sigma_B^2 = 22$. To test whether the average distance differs between the two batteries, $n_A = 9$ vehicles using battery (A) and $n_B = 11$ vehicles using battery (B) were randomly selected, giving distances X_{Ai} $(i = 1, \ldots, 9)$ and X_{Bj} $(j = 1, \ldots, 11)$. Conduct a hypothesis test on the difference of two means at significance level $\alpha = 0.05$.

(a) State the null and alternative hypotheses.

(b) Find the null distribution of the test statistic

$$Z = \frac{\overline{X}_A - \overline{X}_B - 0}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}},$$

where $\overline{X}_A$ is the sample mean for the new battery and $\overline{X}_B$ is the sample mean for the existing battery.

(c) Find the rejection region at significance level $\alpha = 0.05$.

(d) The following summary statistics were obtained. Find the observed value of the test statistic in part (b).

$$\overline{x}_A = \frac{1}{9}\sum_{i=1}^{9} x_{Ai} = 455, \qquad \overline{x}_B = \frac{1}{11}\sum_{j=1}^{11} x_{Bj} = 449.5455$$

(e) Using the results in parts (c) and (d), complete the hypothesis test. State the conclusion in the context of the data.

2. A food scientist wants to test whether Brand 1 wine has a lower mean pH level than Brand 2 wine. To test this at the significance level $\alpha = 0.05$, independent random samples of size $n_1 = 50$ and $n_2 = 50$ are selected from Brand 1 and Brand 2, respectively. Assume the pH levels follow

$$X_{11}, \cdots, X_{1,50} \sim N(\mu_1, 0.01), \qquad X_{21}, \cdots, X_{2,50} \sim N(\mu_2, 0.01),$$

where μ_1 and μ_2 denote the mean pH levels of Brand 1 and Brand 2, respectively.

(a) State the null and alternative hypotheses. (Use a one-sided test.)

(b) Find the test statistic and its null distribution.

(c) Find the rejection region.

(d) Given $\overline{x}_1 = 3.4$ and $\overline{x}_2 = 3.5$, complete the hypothesis test, and state the conclusion in the context of the data.

(e) Find the power of the test assuming $\mu_1 - \mu_2 = -0.05$.

3. A company wants to determine whether the mean signal strength (in dBm) measured from two antennas, A and B, is different. Signal strength was measured $n_A = 22$ times using antenna A and $n_B = 26$ times using antenna B. Let μ_A and μ_B denote the population mean signal strengths. Conduct a two-sample t-test at significance level $\alpha = 0.01$.

(a) State the null and alternative hypotheses. (Use a two-sided test.)

(b) Find the test statistic and its null distribution.

(c) Using the following R code and output, report the p-value and complete the hypothesis test. State the conclusion in the context of the data.

```
Adat <- c(-98.70, -81.70, -83.00, # ... remaining values omitted ...
          -96.20)
Bdat <- c(-84.00, -76.21, -83.50, # ... remaining values omitted ...
          -79.10)
t.test(Adat, Bdat, alternative = "two.sided", var.equal = FALSE, mu = 0)

  Welch Two Sample t-test

data:  Adat and Bdat
t = -0.066942, df = 29.2, p-value = 0.9472
alternative hypothesis: true difference in means is not equal to 0
```

8.6 Difference of two means: equal variances

If there is good reason to believe the two populations share a common variance, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we can estimate that single σ^2 from both samples at once. Then $\mathrm{Var}(\overline{X}_1 - \overline{X}_2) = \sigma^2(1/n_1 + 1/n_2)$, and the **pooled sample variance**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

combines the two estimates, giving $SE(\overline{X}_1 - \overline{X}_2) = S_p\sqrt{1/n_1 + 1/n_2}$.

TWO-SAMPLE t-TEST (EQUAL VARIANCES)

When $\sigma_1^2 = \sigma_2^2$, for $H_0 : \mu_1 - \mu_2 = \delta_0$,

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} \overset{H_0}{\sim} t(n_1 + n_2 - 2), \qquad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

NOTE

The degrees of freedom $n_1 + n_2 - 2$, and the statistic itself, come from the same construction as the one-sample t in Section 8.1.2. Suppose both samples are normal with the common variance σ^2. Standardizing the difference of means with the *known* σ gives a standard normal,

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1).$$

Each sample variance, by the result of Chapter 7, contributes a chi-squared, and because the two samples are independent so are the two:

$$V_1 = \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1), \qquad V_2 = \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1).$$

By the additivity of the chi-squared distribution their sum is again chi-squared; rewriting the numerator with the pooled variance shows it carries exactly $n_1 + n_2 - 2$ degrees of freedom:

$$V = V_1 + V_2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)\,S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

The definition of the t-distribution (Definition 8.1) now assembles these pieces. Forming $Z/\sqrt{V/(n_1 + n_2 - 2)}$, the unknown σ cancels between the numerator and the denominator, leaving the pooled estimate S_p in its place:

$$\frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2).$$

Under H_0 the mean difference $\mu_1 - \mu_2$ equals δ_0, which gives the test statistic above.

EXAMPLE 8.17 • Comparing two user cohorts

Two independent groups of sessions are rated on a quality scale: 100 sessions with a feature enabled and 50 with it disabled. Careful analysis suggests the two groups have the same variance, so we pool.

Group	n	$\bar{x}$	s
Enabled	100	7.18	1.60
Disabled	50	6.78	1.43

Testing $H_0 : \mu_{\text{on}} - \mu_{\text{off}} = 0$ against $H_A : \mu_{\text{on}} - \mu_{\text{off}} > 0$, the pooled variance and standard error are

$$s_p^2 = \frac{99(1.60^2) + 49(1.43^2)}{148} = 2.39, \qquad SE = \sqrt{2.39}\,\sqrt{\tfrac{1}{100} + \tfrac{1}{50}} = 0.27,$$

so $t = (7.18 - 6.78)/0.27 = 1.48$ on $df = 148$. The upper critical value is $t_{0.05}(148) = 1.655$; since $1.48 < 1.655$ (p-value ≈ 0.071), we do *not* reject H_0. This sample does not establish a difference. A non-significant result does not show that no difference exists — only that, at this sample size, the data cannot distinguish zero from a modest positive effect; reporting the confidence interval for the difference makes the range of compatible values explicit (Section 8.3 treats the related question of sensitivity in advance). The pooled (equal-variance) test in R:

```
> t.test(on, off, alternative = "greater", mu = 0, var.equal = TRUE)

 Two Sample t-test

data: on and off
t = 1.4824, df = 148, p-value = 0.07050
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.046407 Inf
sample estimates:
mean of x mean of y
 7.1795 6.7790
```

! CAUTION

Pool only when the equal-variance assumption is genuinely justified. If the two spreads differ noticeably — as in Example 8.15 — the unequal-variance (Welch) test of Section 8.5 is the safer default and is what most software uses unless told otherwise.

As with one mean, a confidence interval for the difference complements the test and is read off the same pieces.

EXAMPLE 8.18 • Interval for the cohort difference

Using the pooled analysis of Example 8.17 ($\bar{x}_{\text{on}} - \bar{x}_{\text{off}} = 0.40$, $SE = 0.27$, $df = 148$), a two-sided

95% interval for $\mu_{\text{on}} - \mu_{\text{off}}$ uses $t_{0.025}(148) = 1.976$:

$$0.40 \pm 1.976(0.27) \;\rightarrow\; (-0.13,\ 0.93).$$

The interval *contains* 0, the same verdict as the test that did not reject H_0: the data are consistent with no difference between the cohorts. Its width also quantifies what the study can rule out — a true difference much beyond about 0.9 on the quality scale would be implausible given these data.

GUIDED PRACTICE 8.7

Two decoding methods are each timed on five independent requests, and their variances are believed equal:

Method A	80	82	84	83	81
Method B	76	78	80	79	77

Test $H_0 : \mu_A - \mu_B = 0$ against $H_A : \mu_A - \mu_B \neq 0$ at $\alpha = 0.05$, given $t_{0.025}(8) = 2.31$.
Answer: $\bar{x}_A = 82, \bar{x}_B = 78$, and each group has $s^2 = 2.5$, so $s_p^2 = 2.5$ and $SE = \sqrt{2.5}\sqrt{1/5 + 1/5} = 1.0$ on $df = 8$. Then $t = (82 - 78)/1.0 = 4.0$; since $4.0 > 2.31$ (p-value ≈ 0.004), reject H_0: the two methods differ in mean time.

EXERCISES (§8.6)

1. In region G, corn yields from fertilizers A and B are normally distributed, with means μ_A and μ_B and equal variances ($\sigma_A^2 = \sigma_B^2 = \sigma^2$). Recently, some farmers have claimed that the mean yields μ_A and μ_B are different. To investigate, researchers selected 12 locations with similar soil quality, randomly applying fertilizer A to 6 sites and fertilizer B to the other 6 sites. Based on the corn yields (unit: kg) from the treated soils, we conduct a hypothesis test for two means at the significance level $\alpha = 0.05$.

   (a) State the null and alternative hypotheses.

   (b) Find the test statistic and its null distribution.

   (c) 12 corn yields (kg) were measured after applying fertilizers A and B. Using the R code and R output provided, write down the p-value and complete the hypothesis test. State the conclusion in the context of the data.

```
fertA <- c(580, 600, 600, 550, 650, 620)
fertB <- c(510, 650, 570, 670, 630, 510)
t.test(fertA, fertB, alternative = "two.sided", mu = 0, var.equal = TRUE)

  Two Sample t-test

data: fertA and fertB
t = 0.31311, df = 10, p-value = 0.7606
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-61.16105 81.16105
sample estimates:
```

```
mean of x mean of y
600 590
```

2. A researcher wants to know whether adults who drink soda every day consume different amounts of sugar compared to adults who rarely drink soda. Daily sugar intake (grams per day) was recorded for $n_E = 12$ everyday soda drinkers and $n_R = 10$ rare soda drinkers. Conduct a two-sample t-test at significance level $\alpha = 0.05$ to determine whether the population mean daily sugar intake of everyday soda drinkers (μ_E) differs from that of rare soda drinkers (μ_R). Assume the variances are equal for both groups ($\sigma_E^2 = \sigma_R^2 = \sigma^2$).

   (a) State the null and alternative hypotheses. (Use a two-sided test.)

   (b) Find the test statistic and its null distribution.

   (c) Using the R code and partial output provided, report the p-value and complete the hypothesis test. State the conclusion in the context of the data.

   (d) Let $\overline{X}_E$ and $\overline{X}_R$ be the sample means for the two groups. Suppose a random variable V satisfies

$$\frac{\overline{X}_E - \overline{X}_R - (\mu_E - \mu_R)}{S_p\sqrt{\frac{1}{n_E} + \frac{1}{n_R}}} = \frac{\overline{X}_E - \overline{X}_R - (\mu_E - \mu_R)}{\sigma\sqrt{\frac{1}{n_E} + \frac{1}{n_R}}} \times \left(\frac{V}{n_E + n_R - 2}\right)^{-1/2},$$

   where S_p is the pooled sample standard deviation. Simplify V and find the distribution of V.

```
everyday <- c(64, 70, 59, 68, 61, 60, 66, 72, 58, 63, 69, 65)
rare     <- c(58, 65, 55, 62, 59, 56, 60, 57, 61, 54)
t.test(everyday, rare, alternative = "two.sided", mu = 0, var.equal = TRUE)

  Two Sample t-test

data:  everyday and rare
t = 3.3673, df = [    ], p-value = [    ]
...(omitted)...
```

8.7 Power calculations for a difference of means

Power for a two-sample test follows the same three-step logic as Section 8.3: find the rejection region in the units of the statistic — here the difference of sample means — suppose a specific true difference, and compute the chance of landing in that region.

EXAMPLE 8.19 • Power of a model A/B test

A new model (treatment) is compared against the current one (control) on mean latency. Earlier studies fix the standard deviation at a *known* $\sigma = 12$ ms in both groups, the test runs at $\alpha = 0.05$, and each group has $n = 100$ requests. We test $H_0 : \mu_{\text{trmt}} - \mu_{\text{ctrl}} = 0$ against the two-sided $H_A : \mu_{\text{trmt}} - \mu_{\text{ctrl}} \neq 0$.

Step 1 — rejection region. The difference of independent sample means is normal, with

$$SE = \sqrt{\frac{\sigma^2}{n_{\text{trmt}}} + \frac{\sigma^2}{n_{\text{ctrl}}}} = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70 \text{ ms}, \qquad \overline{X}_{\text{trmt}} - \overline{X}_{\text{ctrl}} \overset{H_0}{\sim} N(0, 1.70^2).$$

A two-sided test rejects when the difference is large in magnitude, $|\bar{x}_{\text{trmt}} - \bar{x}_{\text{ctrl}}| > c$, with

$$c = z_{0.025} \times SE = 1.96 \times 1.70 \approx 3.33 \text{ ms}$$

(Figure 8.11).
Step 2 — suppose a true difference. If the new model truly lowers latency by 3 ms, the true difference is -3, and $\bar{X}_{\text{trmt}} - \bar{X}_{\text{ctrl}} \overset{H_A}{\sim} N(-3, 1.70^2)$: the sampling distribution shifts left by 3 while its spread is unchanged.
Step 3 — probability of rejecting.

$$\begin{aligned}\text{Power} &= \mathbb{P}\left(|\bar{X}_{\text{trmt}} - \bar{X}_{\text{ctrl}}| > 3.33 \;\middle|\; \mu_{\text{diff}} = -3\right) \\ &= \mathbb{P}\left(Z < \tfrac{-3.33-(-3)}{1.70}\right) + \mathbb{P}\left(Z > \tfrac{3.33-(-3)}{1.70}\right) \\ &= \mathbb{P}(Z < -0.20) + \mathbb{P}(Z > 3.72) \;\approx\; 0.42.\end{aligned}$$

With 100 per group the test has only about a 42% chance of detecting a true 3 ms difference — almost all of it from the lower rejection tail, since the upper tail is negligible once the distribution has shifted left.

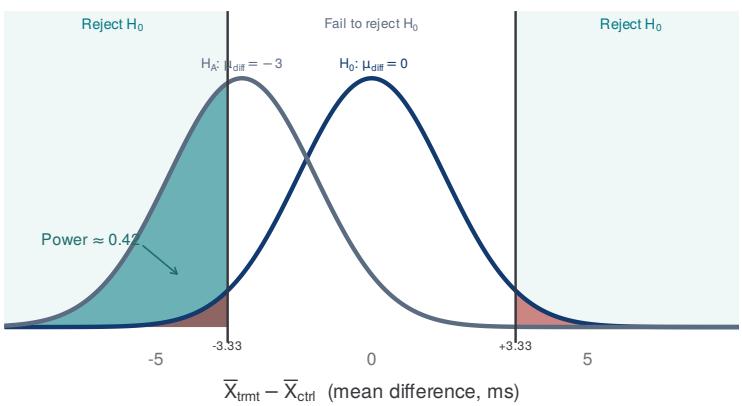


Figure 8.11. Two-sided power for the difference of means. The cutoffs ± 3.33 split the axis into a central "fail to reject" region and two rejection regions. The power is the shaded area of the alternative curve ($\mu_{\text{diff}} = -3$) inside the rejection regions; the two small red slivers of the null curve there total α.

Because power grows with the sample size, we can ask how large n must be to reach a target power. Holding the effect at 3 ms and $\sigma = 12$, the power rises with n and crosses the conventional 80% mark near $n \approx 252$ per group (Figure 8.12). Such curves are the basis of **sample-size planning**: pick the smallest n whose power reaches the desired level.

GUIDED PRACTICE 8.8

For the two-sided A/B test above ($\sigma = 12$ ms, $\alpha = 0.05$), how many requests *per group* give 80% power to detect a $\Delta = 3$ ms difference? Use the two-sample size formula $n \geq 2\sigma^2(z_{\alpha/2} + z_{\beta})^2/\Delta^2$

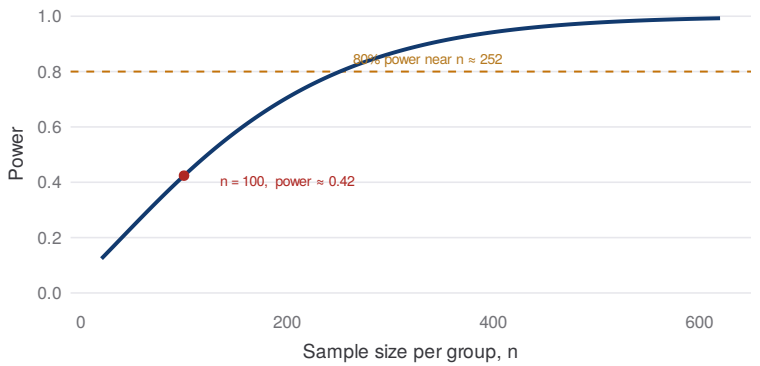


Figure 8.12. Power against sample size per group for detecting a 3 ms difference with $\sigma = 12$. At $n = 100$ the power is about 0.42; reaching 80% power requires roughly 252 per group.

with $z_{0.025} = 1.96$ and $z_{0.20} = 0.842$.
Answer: $n \geq 2(12^2)(1.96 + 0.842)^2/3^2 = 288\,(2.802)^2/9 = 251.2$, so $n = 252$ per group — the value the power curve in Figure 8.12 crosses 80% at.

EXERCISES (§8.7)

1. A health researcher wants to test whether people who exercise regularly sleep more on average than people who do not exercise regularly. Group 1 consists of people who exercise regularly, while Group 2 consists of people who do not exercise regularly. To test this at the significance level $\alpha = 0.05$, the researcher samples $n_1 = 8$ people from Group 1 and $n_2 = 18$ people from Group 2. Assume the daily sleep durations (in hours) follow:

$$X_{11}, \cdots, X_{1,8} \sim N(\mu_1, 1), \qquad X_{21}, \cdots, X_{2,18} \sim N(\mu_2, 2.25),$$

where μ_1 and μ_2 denote the mean daily sleep duration for Group 1 and Group 2, respectively.

   (a) State the null and alternative hypotheses. (Use a one-sided test.)

   (b) Find the test statistic and its null distribution.

   (c) Find the rejection region.

   (d) Given $\bar{x}_1 = 8.2$ and $\bar{x}_2 = 7.0$, complete the hypothesis test. State the conclusion in the context of the data.

   (e) Find the power of the test assuming $\mu_1 - \mu_2 = 1$.

2. A company claims that its new earplugs reduce noise by more than $\delta_0 = 2$ dB on average compared with the existing earplugs. To test this, $n_A = 40$ new and $n_B = 50$ existing earplugs are randomly selected, and their noise reductions are measured. Test at significance level $\alpha = 0.01$. Assume the noise reductions of the new and existing earplugs follow normal distributions with means μ_A, μ_B and standard deviations $\sigma_A = 0.5, \sigma_B = 0.6$.

   (a) State the null and alternative hypotheses.

   (b) $\overline{X}_A - \overline{X}_B$ is used as the test statistic, where $\overline{X}_A$ and $\overline{X}_B$ are the sample means for the new and existing earplugs. Find the null distribution of this test statistic.

   (c) Find the rejection region at significance level $\alpha = 0.01$.

(d) The sample mean noise reductions are $\bar{x}_A = 27.48$ and $\bar{x}_B = 25.17$. Find the observed value of the test statistic and test the hypothesis. State the conclusion in the context of the data.

(e) When the true difference of the two population means is $\mu_A - \mu_B = 2.5$, find the power of the test.

8.8 Comparing many means: ANOVA

The two-sample test compares exactly two means. To compare three or more at once — the average score across several classes, or a metric across several model families — we could run a separate t-test on every pair, but each test carries its own Type 1 risk, and the chance of at least one false alarm grows quickly with the number of pairs. The **analysis of variance** (ANOVA) instead asks a single question,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \qquad H_A : \text{at least one mean differs},$$

and answers it by comparing the variability *between* the group means to the variability *within* the groups.

8.8.1 The F-statistic

If the between-group spread is large relative to the within-group spread, the group means are too far apart to be chance. ANOVA measures each piece with a mean square. The **mean square between groups** captures how far the group means $\bar{x}_i$ sit from the grand mean $\bar{x}$,

$$\mathrm{MSG} = \frac{\mathrm{SSG}}{df_G}, \qquad \mathrm{SSG} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2, \qquad df_G = k - 1,$$

and the **mean square error** pools the within-group variances,

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{df_E}, \qquad \mathrm{SSE} = \sum_{i=1}^{k} (n_i - 1)s_i^2, \qquad df_E = n - k,$$

where $n = n_1 + \cdots + n_k$ is the total sample size.

THE ANOVA F-TEST

For $H_0 : \mu_1 = \cdots = \mu_k$, the test statistic and its null distribution are

$$F = \frac{\mathrm{MSG}}{\mathrm{MSE}} = \frac{\mathrm{SSG}/df_G}{\mathrm{SSE}/df_E} \overset{H_0}{\sim} F(k - 1, n - k),$$

and large values of F are evidence against H_0. The p-value is the upper-tail area $\mathbb{P}(F \geq F_{\mathrm{obs}})$.

NOTE

The three sums of squares fit together. Writing x_{ij} for the j-th observation in group i and splitting each deviation from the grand mean as $(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})$, the cross term sums to zero, leaving the **total sum of squares** as the sum of the within- and between-group parts:

$$\underbrace{\sum_i \sum_j (x_{ij} - \bar{x})^2}_{\text{SST} = (n-1)s^2} = \underbrace{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}_{\text{SSE}} + \underbrace{\sum_i \sum_j (\bar{x}_i - \bar{x})^2}_{\text{SSG}}.$$

The reference distribution is new. Just as a ratio of a normal to a chi-squared gave the t, a ratio of two independent chi-squareds gives the F.

DEFINITION 8.8 • The F-distribution

If $V_1 \sim \chi^2(k_1)$ and $V_2 \sim \chi^2(k_2)$ are independent, then

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2),$$

the F-distribution with k_1 and k_2 degrees of freedom.

NOTE

The F-statistic matches this definition. Model the data as $x_{ij} = \mu_i + e_{ij}$ with $e_{ij} \sim N(0, \sigma^2)$ independent. From the chi-squared theory of Chapter 7, the pooled within-group variation gives $V_2 = \text{SSE}/\sigma^2 \sim \chi^2(n-k)$, and *when* H_0 *holds* the between-group variation gives $V_1 = \text{SSG}/\sigma^2 \sim \chi^2(k-1)$, independently of V_2. Hence

$$F = \frac{V_1/(k-1)}{V_2/(n-k)} = \frac{\text{SSG}/(k-1)}{\text{SSE}/(n-k)} = \frac{\text{MSG}}{\text{MSE}} \overset{H_0}{\sim} F(k-1,\, n-k).$$

With only two groups this reduces to the equal-variance t-test: if $T \sim t(k)$ then $T^2 \sim F(1, k)$, because $T = Z/\sqrt{V/k}$ with $Z^2 \sim \chi^2(1)$ gives $T^2 = (Z^2/1)/(V/k)$.

EXAMPLE 8.20 • Success rate across model families

Three families of language models — A, B, and C — are evaluated on a hard reasoning benchmark, and each model's success rate is recorded.

	Family A	Family B	Family C	All
Sample size (n_i)	160	205	64	429
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302	0.317
Sample SD (s_i)	0.043	0.038	0.038	0.040

With $k = 3$ groups and $n = 429$, the between- and within-group sums of squares are

$$\text{SSG} = \sum_i n_i(\bar{x}_i - \bar{x})^2 = 0.0161, \qquad \text{SSE} = \sum_i (n_i - 1)s_i^2 = 0.6740,$$

so $df_G = k - 1 = 2$ and $df_E = n - k = 426$, giving

$$\text{MSG} = \frac{0.0161}{2} = 0.00803, \quad \text{MSE} = \frac{0.6740}{426} = 0.00158, \quad F = \frac{0.00803}{0.00158} = 5.08.$$

The p-value is the upper-tail area of an $F(2, 426)$ beyond 5.08 (Figure 8.13):

```
> pf(5.08, df1 = 2, df2 = 426, lower.tail = FALSE)
[1] 0.006602098
```

Since $0.0066 < 0.05$, we reject H_0: mean success rate differs across the families. The same fit and ANOVA table in R:

```
> mod <- lm(success ~ family, data = models)
> anova(mod)
Analysis of Variance Table

Response: success
 Df Sum Sq Mean Sq F value Pr(>F)
family 2 0.01606 0.0080314 5.0766 0.006624 **
Residuals 426 0.67395 0.0015820
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

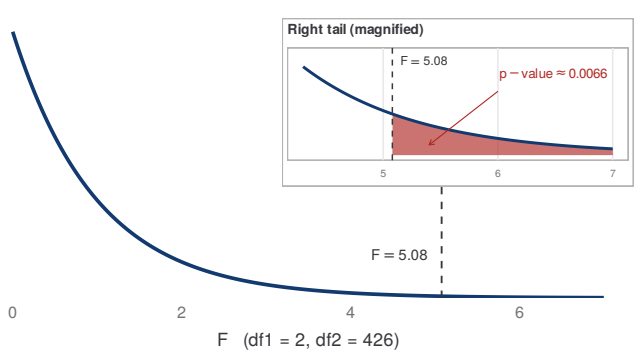


Figure 8.13. The $F(2, 426)$ null distribution. The observed $F = 5.08$ sits far in the right tail, leaving an area of about 0.0066 beyond it.

GUIDED PRACTICE 8.9

Three prompting strategies are each tried on four tasks, with scores

Strategy	Scores			
Zero-shot	68	76	70	74
Few-shot	74	82	76	80
Chain-of-thought	80	88	82	86

The group means are 72, 78, 84 and the grand mean is 78. Build the ANOVA table and test equality of the three mean scores at $\alpha = 0.05$, given $F_{0.05}(2, 9) = 4.26$.
Answer: each group has within-group sum of squares 40, so SSE $= 3(40) = 120$ on $df_E = 9$ and MSE $= 13.33$. The between-group sum of squares is SSG $= 4\,[(72-78)^2+(78-78)^2+(84-78)^2] = 288$ on $df_G = 2$, so MSG $= 144$ and $F = 144/13.33 = 10.8$. Since $10.8 > 4.26$ (p-value ≈ 0.004), reject H_0: mean score differs across the three strategies.

8.8.2 After a significant F: pairwise comparisons

A significant F reports only that the means are not all equal; it does not say *which* differ. The natural follow-up compares the groups in pairs, but running several tests revives the multiple-testing problem, so each comparison is held to the stricter Bonferroni level $\alpha^\star = \alpha/K$ for K comparisons. Each pairwise statistic reuses the pooled within-group spread MSE from the ANOVA as its variance estimate, on df_E degrees of freedom,

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{MSE}\left(1/n_i + 1/n_j\right)}} \sim t(df_E).$$

EXAMPLE 8.21 • Which model families differ?

The F-test of Example 8.20 rejected equality of the three families' mean success rates, with MSE $= 0.00158$ on $df_E = 426$. There are $K = \binom{3}{2} = 3$ pairs, so the Bonferroni level is $\alpha^\star = 0.05/3 = 0.0167$. Using $\bar{x}_A = 0.320$, $\bar{x}_B = 0.318$, $\bar{x}_C = 0.302$ with $n_A = 160$, $n_B = 205$, $n_C = 64$,

Pair	$\bar{x}_i - \bar{x}_j$	SE	t	p-value
A vs B	0.002	0.0042	0.48	0.634
A vs C	0.018	0.0059	3.06	0.0024
B vs C	0.016	0.0057	2.81	0.0052

Comparing each p-value to $\alpha^\star = 0.0167$, family C differs significantly from both A and B, while A and B are statistically indistinguishable. The single significant F thus resolves into a concrete statement: family C, the lowest performer, is the one that stands apart.

8.8.3 Checking conditions

The F-test rests on three conditions, to be checked before trusting its p-value: the observations are **independent** within and between groups; the data in each group are **approximately normal** (most

important for small groups); and the groups have roughly **constant variance** (most important when the group sizes differ).

> **NOTE**
>
> A significant F says only that *some* pair of means differs, not which. Following up with a t-test on every pair reintroduces the multiple-testing problem ANOVA was meant to avoid, so the pairwise tests should use a stricter level — for example the Bonferroni level $\alpha^\star = \alpha/K$ for $K = \binom{k}{2}$ comparisons.

EXERCISES (§8.8)

1. State the ANOVA hypotheses for comparing the mean score across four lectures, and give the degrees of freedom df_G and df_E if each lecture has 50 students.
2. An ANOVA reports SSG = 1290 on $df_G = 2$ and SSE = 29810 on $df_E = 161$. Compute MSG, MSE, and F.
3. With $k = 4$ groups, how many pairwise comparisons are there, and what is the Bonferroni level $\alpha^\star$ for $\alpha = 0.05$?
4. A survey on cadet life satisfaction was conducted during a training session at a military academy. A total of 268 responses were grouped by company (8 companies in total). Conduct a one-way ANOVA to test whether the average cadet life satisfaction score differs across companies at significance level $\alpha = 0.05$. Assume that all conditions for ANOVA are satisfied.

Company	1	2	3	4	5	6	7	8	All
Mean	3.80	3.45	3.68	3.22	3.91	3.34	3.70	3.51	3.58
SD	0.25	0.28	0.22	0.27	0.20	0.30	0.24	0.29	0.27
n	33	35	30	32	34	35	32	37	268

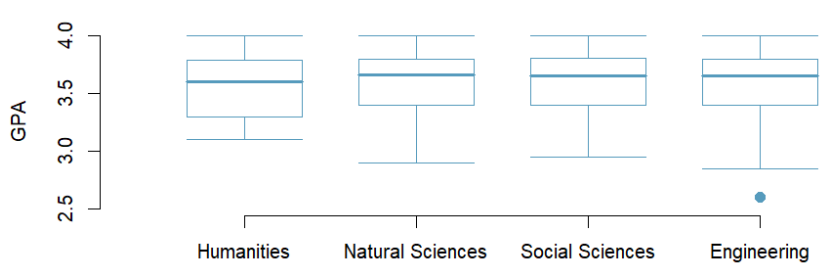
   (a) Let $\mu_1, \mu_2, \ldots, \mu_8$ represent the true mean cadet life satisfaction for each of the 8 companies. State the null and alternative hypotheses.

   (b) Below is part of the ANOVA output. Fill in the empty cells.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Company		12.9853			< 0.001
Residuals		17.5223			
Total	267	30.5076			

   (c) Do you reject the null hypothesis? State the conclusion in the context of the data.

5. A total of 268 undergraduate students taking an introductory statistics course conducted a survey on GPA and major. We conduct a one-way ANOVA to determine whether the average GPA differs across college majors at significance level $\alpha = 0.05$. The side-by-side boxplots show the distribution of GPA among four groups of majors, and the summary statistics and ANOVA output are provided below. (Assume that all conditions for conducting ANOVA are satisfied.)



College Major	Humanities	Natural Sciences	Social Sciences	Engineering	All
Mean	3.58	3.61	3.57	3.59	3.59
SD	0.30	0.27	0.26	0.27	0.28
n	113	52	33	70	268

(a) State the null and alternative hypotheses. (Define the parameters of interest.)

(b) Below is part of the output associated with this test. Fill in the empty cells.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Major		0.0348			0.931
Residuals		20.6831			
Total	267	20.7179			

(c) What is the conclusion of the hypothesis test? State the conclusion in the context of the data.

8.9 Summary of tests for means

The tests of this chapter share one template — a statistic of the form (estimate − null value)/(standard error), compared against a t, normal, or F reference. The table collects them.

	Null hypothesis	Test statistic	df
One-sample t-test	$H_0 : \mu = \mu_0$	$T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(df)$	$n - 1$
One-sample Z-test (known variance)	$H_0 : \mu = \mu_0$	$Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	—
Paired t-test	$H_0 : \mu_{\text{diff}} = \delta_0$	$T = \dfrac{\overline{X}_{\text{diff}} - \delta_0}{S_{\text{diff}}/\sqrt{n_{\text{diff}}}} \sim t(df)$	$n_{\text{diff}} - 1$
Two-sample t-test	$H_0 : \mu_1 - \mu_2 = \delta_0$	$T = \dfrac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(df)$	ψ
Two-sample t-test (equal variances)	$H_0 : \mu_1 - \mu_2 = \delta_0$	$T = \dfrac{\overline{X}_1 - \overline{X}_2 - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(df)$	$n_1 + n_2 - 2$
Two-sample Z-test (known variances)	$H_0 : \mu_1 - \mu_2 = \delta_0$	$Z = \dfrac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$	—
ANOVA F-test	$H_0 : \mu_1 = \cdots = \mu_k$	$F = \dfrac{\text{MSG}}{\text{MSE}} \sim F(k-1, n-k)$	$(k-1, n-k)$

Choosing among them is largely mechanical once two questions are answered: *how many groups*, and *how are they related*? One group is a one-sample test; two groups call for a paired test when the observations are matched and a two-sample test when they are independent; three or more groups call for ANOVA. Within the two-sample case, equal variances permit pooling, while unequal or unknown variances call for the Welch form — the safe default. Whether σ is known decides z versus t, though in practice σ is almost never known and the t-based rows are the ones used.

EXAMPLE 8.22 • Choosing the right test

Four short scenarios, each resolved by the two questions above.

1. *The same* 40 *documents are summarized by two models, and the mean summary lengths are compared.* Two groups, matched on the document — a **paired** t**-test** on the per-document differences.
2. *Mean accuracy is compared across five different model architectures.* More than two groups — an **ANOVA** F**-test**, followed by pairwise comparisons if it is significant.
3. *The mean latency of* 50 *requests is tested against a published target of* 200 *ms, with* σ *unknown.* One group against a fixed value — a **one-sample** t**-test**.
4. *Throughput is compared between two independent server pools whose sample SDs differ noticeably.* Two independent groups with unequal spreads — a **two-sample (Welch)** t**-test**.

In every case the statistic is the same (estimate − null value) over standard error; only the standard error and the reference distribution change.

CHAPTER 9

Linear Regression

Every inference method so far has examined a single quantity — one mean, one proportion, or a difference between two of them. This chapter turns to the relationship *between* two numerical variables: how a **response** y moves with a **predictor** x. The tool is **linear regression**, which summarizes that relationship with a straight line and uses it to predict y from x. Section 9.1 fits a line, measures how far the data fall from it with **residuals**, and quantifies the strength of the linear association through the **correlation coefficient**. Section 9.2 makes "best fit" precise with the **least squares** criterion and derives closed-form estimates of the slope and intercept. Section 9.3 carries the t-based machinery of Chapter 8 into regression: a hypothesis test and a confidence interval for the slope, together with the conditions that make them trustworthy. Finally, Section 9.4 moves from one predictor to many. It introduces **multiple linear regression**, folds categorical predictors in through **indicator variables**, warns about **collinearity** among correlated predictors, and uses **adjusted** R^2 to compare models of different sizes.

9.1 Line fitting, residuals, and correlation

9.1.1 Fitting a line to data

A few relationships are exactly linear. A cloud provider that charges a flat monthly fee plus a fixed rate per GPU-hour produces invoices that lie perfectly on a line (Figure 9.1): the hours used determine the bill with no error.

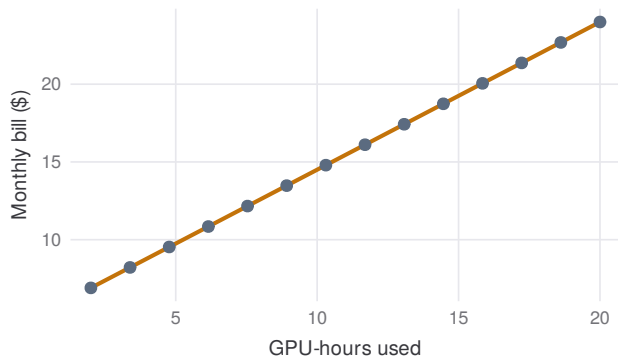


Figure 9.1. An exact linear relationship: each month's bill is a fixed fee plus a constant rate per GPU-hour, so every point lies on one line.

Real data are never this tidy, so we model the relationship with a line *plus* a random error that absorbs everything the line misses.

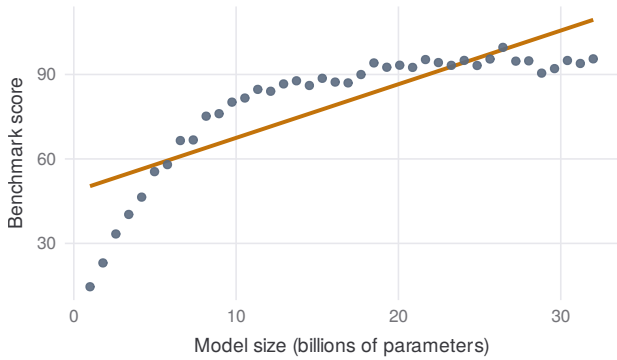


Figure 9.2. A saturating relationship. The straight line is a poor description: it overshoots in the middle and undershoots at both ends.

DEFINITION 9.1 • Linear regression model

The **linear regression model** relates a predictor x to a response Y by

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where the **regression coefficients** β_0 (intercept) and β_1 (slope) are fixed unknown parameters and ε is a random **error**. The coefficients are estimated from data; the estimates are written $\hat{\beta}_0$ and $\hat{\beta}_1$, giving the fitted line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Here x is the **predictor** (or explanatory) variable and y is the **response**; we use x to predict y, not the reverse. Most data form a cloud of points scattered around a trend rather than a perfect line, and a line is only worth fitting when that trend is roughly straight. When the pattern bends — as when a model's benchmark score climbs quickly with size and then saturates (Figure 9.2) — a straight line systematically misses the shape of the data, and a linear fit is the wrong summary.

9.1.2 The regression equation and prediction

EXAMPLE 9.1 • Predicting held-out performance

A team evaluates 104 language models on two suites: a *public* benchmark whose questions circulate openly, and a *private* benchmark held out to detect overfitting. Figure 9.3 plots each model's private score (y) against its public score (x); the two are positively associated, so a line is a reasonable summary. Fitting the model (Section 9.2 shows how) gives

$$\hat{y} = 41 + 0.59\,x.$$

Predict the held-out score of a model that scores 80 publicly. Substituting $x = 80$,

$$\hat{y} = 41 + 0.59 \times 80 = 88.2.$$

A model scoring 80 publicly is predicted to *average* 88.2 on the private benchmark — 88.2 is

the mean held-out score among all models at $x = 80$, not a guarantee for any single one.

Figure 9.3. Private versus public benchmark score for 104 models, with the fitted line $\hat{y} = 41 + 0.59x$. The association is positive and roughly linear.

9.1.3 Residuals

A fitted line captures the trend, but individual points fall above or below it. Those vertical gaps are the model's leftovers.

DEFINITION 9.2 • Residual

The **residual** of the i-th observation (x_i, y_i) is the difference between the observed and fitted response,

$$e_i = y_i - \hat{y}_i, \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Equivalently, data = fit + residual. A positive residual lies above the line; a negative one lies below.

EXAMPLE 9.2 • Computing a residual

One model scores $x = 85$ publicly and $y = 98.6$ privately. With $\hat{y} = 41 + 0.59x$,

$$\hat{y}_i = 41 + 0.59 \times 85 = 91.15, \qquad e_i = y_i - \hat{y}_i = 98.6 - 91.15 = 7.45.$$

The model beats its predicted held-out score by 7.45 points (Figure 9.4).

Graphing the residuals against x flattens the trend and makes leftover structure easy to see.

Figure 9.4. Residuals are vertical distances from each point to the line. The highlighted model at $(85, 98.6)$ lies 7.45 above its fitted value.

DEFINITION 9.3 • Residual plot

A **residual plot** graphs the residuals e_i against the predictor x_i (or the fitted values $\hat{y}_i$). When the line captures the relationship, the residuals scatter randomly around the dashed line at 0 with no pattern (Figure 9.5).

Figure 9.5. Residual plot for the benchmark fit. The points spread evenly above and below 0 with no curve or funnel, supporting a linear model.

NOTE

A residual plot is a diagnostic, not a verdict. Leftover *curvature* signals that the relationship is nonlinear; a *funnel* shape, with spread growing along x, signals non-constant error variance. Both are easier to spot here than in the original scatterplot, where the slope hides them.

9.1.4 Describing linear association with correlation

The slope reports the direction of a linear relationship but not its *strength*. For that we use a unitless summary.

DEFINITION 9.4 • Sample correlation coefficient

For paired observations $(x_1, y_1), \dots, (x_n, y_n)$, the **sample correlation coefficient** is

$$R = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}},$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

The correlation measures the strength of the *linear* association and always lies in $[-1, 1]$: $R = \pm 1$ is a perfect line, $R = 0$ is no linear association (though variables with $R = 0$ can still be related *nonlinearly*), and its sign matches the slope's. Because it is built from standardized deviations, R is unaffected by rescaling or shifting either variable — measuring prompt length in thousands of tokens, or a score on a different scale, leaves it unchanged. Figure 9.6 shows the visual meaning of several values.

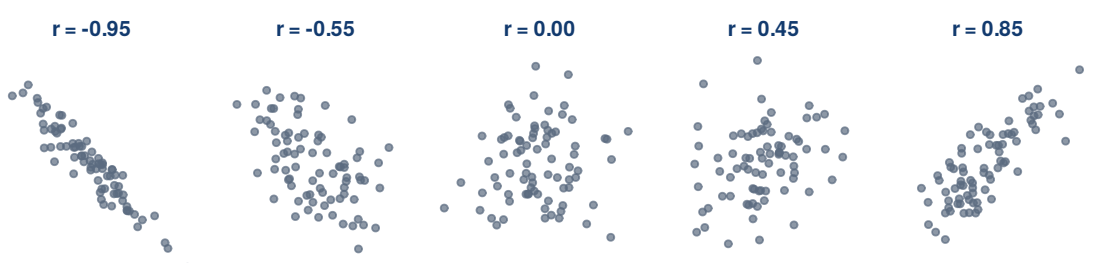


Figure 9.6. Scatterplots of increasing correlation. Strength is how tightly the points hug a line; the sign is the direction of the trend.

For the benchmark data the points follow an upward line fairly tightly but with real scatter, consistent with $R = 0.69$.

In Chapter 4 we defined the *population* correlation of two random variables,

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

The sample correlation R estimates ρ, so $R = 0.69$ gives $\hat{\rho} = 0.69$ for the benchmark population. This lets us test whether any linear relationship exists at all,

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_A : \rho \neq 0.$$

CORRELATION TEST FOR ρ

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent draws from a **bivariate normal** distribution, then under $H_0 : \rho = 0$

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \overset{H_0}{\sim} t(n-2),$$

where R is the sample correlation. A bivariate normal distribution is the two-dimensional generalization of the normal: both margins and every slice through it are normal.

NOTE

Because correlation is invariant to shifting and scaling, assume without loss of generality that the pairs are standardized to mean 0 and standard deviation 1. Under H_0 the pairs are bivariate normal with $\rho = 0$, so X_i and Y_i are independent. Condition on $X_1 = x_1, \ldots, X_n = x_n$. Then

$$Z := \frac{S_{xY}}{\sqrt{S_{xx}}} = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sqrt{\sum_i (x_i - \bar{x})^2}} \sim N(0, 1),$$

and it is known that

$$V := S_{YY} - \frac{S_{xY}^2}{S_{xx}} \sim \chi^2(n-2),$$

independent of Z. Hence, conditional on the x's,

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{Z}{\sqrt{V/(n-2)}} \sim t(n-2).$$

Since this conditional distribution does not depend on the x's, it is also the unconditional null distribution.

EXAMPLE 9.3 • Is the association real?

Test whether public and private benchmark scores are correlated, using $n = 104$ and $r = 0.69$, with $H_0 : \rho = 0$ against $H_A : \rho \neq 0$ and assuming independent models from a bivariate normal population. The observed statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.69\sqrt{102}}{\sqrt{1-0.69^2}} = \frac{6.97}{0.7238} \approx 9.63,$$

on $t(102)$. The two-sided p-value is astronomically small, so we reject H_0: there is overwhelming evidence of a positive linear association.

```
> cor.test(public, private, alternative = "two.sided")

  Pearson's product-moment correlation

data: public and private
t = 9.6569, df = 102, p-value = 4.681e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5750415 0.7798823
sample estimates:
 cor
0.6910937
```

GUIDED PRACTICE 9.1

An edge-AI team measures, for 12 phone models, the on-device model size x (MB) and the battery life y (minutes of continuous inference). Suspecting that larger models drain the battery faster, they test for a *negative* correlation, assuming a bivariate normal population. The summaries are

$$n = 12, \quad S_{xx} = 34.0961, \quad S_{yy} = 2762.25, \quad S_{xy} = -214.735.$$

State the hypotheses, compute r and the statistic, and conclude.

Answer: One-sided, $H_0 : \rho = 0$ versus $H_A : \rho < 0$.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-214.735}{\sqrt{34.0961 \times 2762.25}} = -0.70, \qquad t = \frac{r\sqrt{10}}{\sqrt{1-r^2}} \approx -3.10$$

on $t(10)$. The lower-tail p-value $\mathbb{P}(T \le -3.10) \approx 0.006$ is below 0.05, so reject H_0: larger on-device models are associated with shorter battery life.

EXERCISES (§9.1)

1. We wish to investigate whether there is a positive linear relationship between the summer daily maximum temperature X and a café's delivery sales Y. To do so, $n = 31$ days were randomly sampled, and for each day the maximum temperature and the café's delivery sales were recorded, giving the results below. Assume that (X, Y) follows a bivariate normal distribution.

i	1	2	3	4	5	6	7	$\cdots$	30	31
Max. temperature (x_i, °C)	33.2	32.5	34.2	36.2	33.7	33.5	33.4	$\cdots$	32.9	33
Delivery sales (y_i, 10,000 KRW)	30	25	40	55	20	27	22	$\cdots$	30	40

   (a) State the null and alternative hypotheses.

   (b) Find the null distribution of the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}},$$

where R is the sample correlation coefficient.

(c) Find the rejection region at significance level $\alpha = 0.05$.

(d) The R code and its partial output for this problem are shown below. Use them to find the p-value and complete the hypothesis test at significance level $\alpha = 0.05$. State the conclusion in the context of the data.

```
> temp <- c(33.2, 32.5, 34.2, 36.2, 33.7, 33.5, 33.4, ..., 32.9, 33)
> sales <- c(30, 25, 40, 55, 20, 27, 22, ..., 30, 40)
> cor.test(temp, sales, alternative = "greater")

Pearson's product-moment correlation

data: temp and sales
t = 4.0776, df = [ ], p-value = [ ]
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.3696727 1.0000000
sample estimates:
cor
0.603664
```

2. A sample of $n = 42$ adults was selected from an insurance study. For each individual, their age and annual medical charges (in hundreds of dollars) were recorded. We wish to investigate whether age (x) is associated with annual medical charges (y). Conduct a hypothesis test on the correlation coefficient ρ to determine whether there is a correlation between age and annual medical charges. (Assume that all necessary conditions are satisfied.)

(a) State the null and alternative hypotheses. (Use a two-sided test.)

(b) Find the test statistic and its null distribution.

(c) Compute the sample correlation coefficient r and the observed test statistic using the following summary statistics.

$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$	$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
361	576	319.2

(d) Compute the p-value and complete the hypothesis test.

3. A café headquarters wants to test whether there is a correlation (ρ) between the floor area (X) and the weekly sales (Y) of a branch. A random sample of $n = 10$ branches gives the results below. Test at significance level $\alpha = 0.05$. Assume (X, Y) follows a bivariate normal distribution.

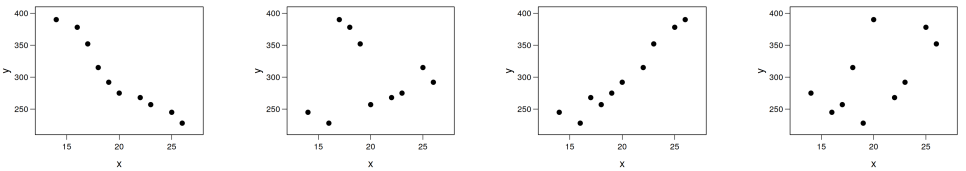
Floor area (x_i, pyeong)	14	19	25	$\cdots$	26	17	20
Weekly sales (y_i, 10,000 KRW)	245	275	378	$\cdots$	390	268	292

(a) State the null and alternative hypotheses.

(b) The test statistic for part (a) is $\dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$, where R is the sample correlation coefficient. Find the null distribution of this test statistic.

(c) Find the rejection region at significance level $\alpha = 0.05$. (Use $t_{0.025}(8) = 2.3060$.)

(d) When the observed sample correlation coefficient is $r = 0.9652$, choose the figure below that best represents the scatter plot of the floor area (X) and the weekly sales (Y).



① ② ③ ④

(e) When the observed sample correlation coefficient is $r = 0.9652$, find the observed value of the test statistic and test the hypothesis using the rejection region in part (c). State the conclusion in the context of the data.

4. A sports league collected data on 11 randomly selected players, measuring their total game time (x, hours) and annual salary (y, USD 1M). We conduct a hypothesis test for the population correlation coefficient ρ to determine whether there is a positive relationship between total game time and annual salary at significance level $\alpha = 0.05$. Assume that game time and salary follow a bivariate normal distribution.

Player (i)	1	2	3	4	5	6	7	8	9	10	11
Game time (x_i)	10	15	20	25	30	12	18	28	22	16	24
Salary (y_i)	2.0	2.2	2.5	3.0	3.5	2.1	2.4	2.9	2.6	3.0	3.2

$$\bar{x} = 20, \quad \bar{y} = 2.6727, \quad S_{xx} = 418, \quad S_{yy} = 2.3418, \quad S_{xy} = 26.6.$$

(a) State the null and alternative hypotheses. (Use a one-sided test.)

(b) Find the null distribution of the test statistic $\dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$, where R is the sample correlation and n is the sample size.

(c) Compute the observed sample correlation coefficient r.

(d) Compute the observed test statistic for part (b).

(e) Compute the p-value and complete the hypothesis test, using $\mathbb{P}(T > 4.8448) \approx 0.0004$ for $T \sim t(9)$. State the conclusion in the context of the data.

9.2 Fitting a line by least squares

Eyeballing a line is unsatisfying: different people draw different lines. We need an objective rule that selects one "best" line from the data.

9.2.1 The least squares criterion

A good line makes the residuals small. Summing the residuals themselves is useless, since positives and negatives cancel, so we sum their *squares* and pick the line that makes the total as small as possible.

DEFINITION 9.5 • Least squares criterion

For data (x_i, y_i) from the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the **least squares** estimates $\hat{\beta}_0, \hat{\beta}_1$ are the candidate intercept and slope (b_0, b_1) that minimize the sum of squared *residuals*

$$S(b_0, b_1) = \sum_{i=1}^{n} \big[\, y_i - (b_0 + b_1 x_i) \,\big]^2 = \sum_{i=1}^{n} e_i(b_0, b_1)^2,$$

where $e_i(b_0, b_1) = y_i - (b_0 + b_1 x_i)$ is the residual of the candidate line. The residuals are computed from the observed data, unlike the model errors ε_i, which are unknown.

Squaring penalizes one large miss far more than several small ones, so the least squares line is pulled toward keeping every point reasonably close.

9.2.2 The least squares estimates

Minimizing S has a clean closed-form solution.

LEAST SQUARES ESTIMATES

The estimates that minimize $S(b_0, b_1)$ are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{s_y}{s_x} R, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where R is the sample correlation, $\bar{x}, \bar{y}$ are the sample means, and s_x, s_y are the sample standard deviations. The fitted line always passes through $(\bar{x}, \bar{y})$.

NOTE

Setting the partial derivatives of S to zero gives the **normal equations**

$$\frac{\partial S}{\partial b_0} = -2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$
$$\frac{\partial S}{\partial b_1} = -2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = 0.$$

The first gives $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, i.e. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Substituting into the second and simplifying,

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \underbrace{\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}}_{s_y/s_x} \cdot \underbrace{\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}}_{R} = \frac{s_y}{s_x} R.$$

EXAMPLE 9.4 • Throughput versus prompt length

An inference platform records, for 50 deployments, the typical prompt length x (tokens) and the sustained throughput y (requests/hour). Longer prompts cost more compute per request, so we expect a negative slope (Figure 9.7). The summaries are

	Prompt length (x)	Throughput (y)
Mean	101,780	19,940
Std. dev.	63,200	5,460
Correlation		$R = -0.499$

Then

$$\hat{\beta}_1 = \frac{s_y}{s_x} R = \frac{5460}{63200} \times (-0.499) = -0.0431,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 19{,}940 - (-0.0431)(101{,}780) = 24{,}319,$$

giving $\hat{y} = 24{,}319 - 0.0431\,x$. Software agrees:

```
> g <- lm(throughput ~ prompt_len, data = serving)
> summary(g)$coefficients
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.4319e+04 1.2915e+03 18.8310 8.2810e-24
prompt_len -4.3072e-02 1.0809e-02 -3.9846 2.2887e-04
```

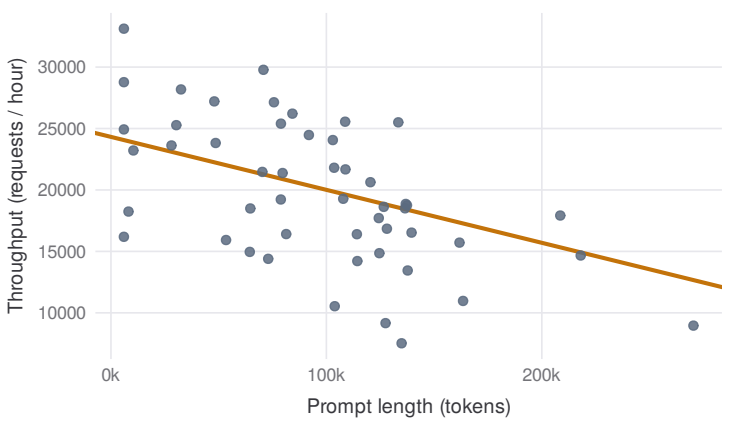


Figure 9.7. Throughput versus prompt length for 50 deployments, with the least squares line $\hat{y} = 24{,}319 - 0.0431x$.

9.2.3 Interpreting the coefficients

The slope and intercept carry concrete meaning in the units of the data. The **slope** $\hat{\beta}_1 = -0.0431$ is the expected change in y per one-unit increase in x: each additional token of prompt length lowers sustained throughput by about 0.0431 requests/hour, roughly 43 requests/hour per extra 1,000 tokens. The **intercept** $\hat{\beta}_0 = 24{,}319$ is the expected y when $x = 0$: a deployment with empty prompts would sustain about 24,319 requests/hour — a reading that is meaningful only when $x = 0$ is sensible and

near the data.

> **! CAUTION**
>
> A regression slope is an *association*, not a cause. Unless the data come from a randomized experiment, $\hat{\beta}_1$ describes how y and x move together, not what would happen if we intervened on x.

9.2.4 Prediction and the danger of extrapolation

Plugging an x into the fitted line gives a **prediction** $\hat{y}$. Inside the range of the observed data this is the line's job. Outside that range — **extrapolation** — the line is a guess about territory the data never visited, and it can fail badly.

> **EXAMPLE 9.5 • An impossible prediction**
>
> Using $\hat{y} = 24{,}319 - 0.0431x$ to predict throughput for a 1,000,000-token prompt gives
>
> $$\hat{y} = 24{,}319 - 0.0431 \times 1{,}000{,}000 = -18{,}781,$$
>
> a *negative* throughput, which is impossible (Figure 9.8). The observed prompts top out near 250,000 tokens; far beyond that the linear trend simply does not hold.

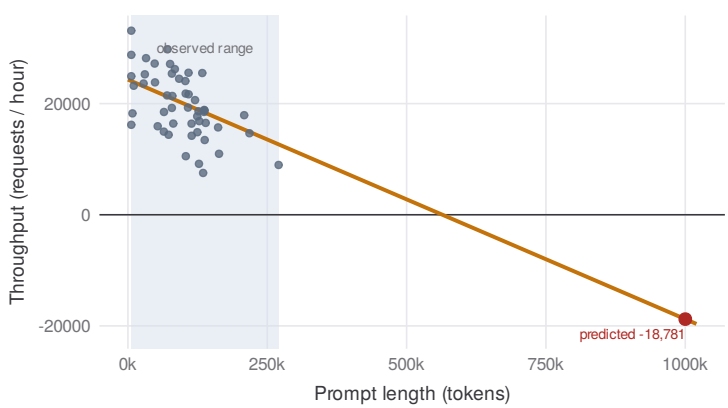


Figure 9.8. The fitted line extended far past the observed range (shaded). At 1,000,000 tokens it predicts a negative throughput — a warning against extrapolation.

9.2.5 Measuring fit with R^2

How much of the variation in y does the line account for? The answer is the square of the correlation.

DEFINITION 9.6 • Coefficient of determination

The **coefficient of determination** R^2 is the square of the sample correlation and equals the proportion of variability in y explained by the regression:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST},$$

where, writing $\hat{y}_i$ for the fitted values,

$$SSR = \sum_i (\hat{y}_i - \overline{y})^2 \qquad \text{(regression sum of squares)},$$
$$SSE = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{(error sum of squares)},$$
$$SST = \sum_i (y_i - \overline{y})^2 = SSR + SSE \qquad \text{(total sum of squares)}.$$

NOTE

From the definition of R,

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{S_{xx}}{S_{yy}}\hat{\beta}_1^2 = \frac{\sum_i (\hat{\beta}_1 x_i - \hat{\beta}_1 \overline{x})^2}{S_{yy}} = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{SSR}{SST}.$$

The decomposition $SST = SSR + SSE$ holds because the cross term vanishes:

$$SST = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = SSE + SSR + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}),$$

and the final sum is zero by the normal equations.

The error variance σ^2 is estimated by spreading SSE over its degrees of freedom.

ESTIMATING THE ERROR VARIANCE

For the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with independent errors of mean 0 and variance σ^2, the **mean squared error**

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

is unbiased for σ^2, that is $E(\hat{\sigma}^2) = \sigma^2$. Two degrees of freedom are lost for the two estimated coefficients.

EXAMPLE 9.6 • Variance explained for throughput

For the serving data the total and error sums of squares are $SST \approx 1.46 \times 10^9$ and $SSE \approx$

1.10×10^9 (in (requests/hour)^2), so

$$R^2 = \frac{SST - SSE}{SST} = \frac{1.46 - 1.10}{1.46} \approx 0.25.$$

Prompt length explains about 25% of the variation in throughput; the rest reflects other factors and noise. The estimated error standard deviation is $\hat{\sigma} = \sqrt{MSE} = \sqrt{SSE/48} \approx 4{,}783$ requests/hour. The full summary collects these quantities:

```
> g <- lm(throughput ~ prompt_len, data = serving)
> summary(g)

Call:
lm(formula = throughput ~ prompt_len, data = serving)

Residuals:
 Min 1Q Median 3Q Max
-10112.8 -3623.4 -216.1 3158.7 11570.7

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.432e+04 1.291e+03 18.831 < 2e-16 ***
prompt_len -4.307e-02 1.081e-02 -3.985 0.000229 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4783 on 48 degrees of freedom
Multiple R-squared: 0.2486, Adjusted R-squared: 0.2329
F-statistic: 15.88 on 1 and 48 DF, p-value: 0.0002289
```

9.2.6 A categorical predictor with two levels

A predictor need not be numerical. With a two-level category, code it as an **indicator** — 1 for one level, 0 for the other — and the regression machinery is unchanged.

EXAMPLE 9.7 • New versus previous-generation GPUs

An operations team models the hourly spot price of a GPU instance from its generation, setting `is_new` = 1 for the current generation and 0 for the previous one:

$$\widehat{\text{price}} = 42.87 + 10.90 \times \texttt{is\_new}.$$

The intercept 42.87 is the average price of a previous-generation instance (where `is_new` = 0); the slope 10.90 is the average price *difference*, so a current-generation instance averages 42.87 + 10.90 = 53.77 dollars/hour (Figure 9.9). With only two levels, the fitted line simply connects the two group means.

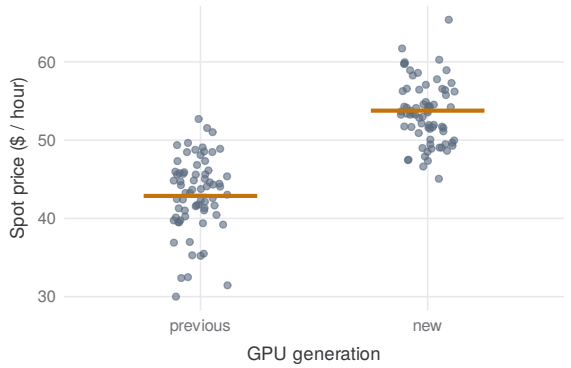


Figure 9.9. Spot price by GPU generation. The fit returns each group's mean: the intercept is the previous-generation average and the slope is the premium for the current generation.

EXERCISES (§9.2)

1. A least squares fit has $\bar{x} = 50$, $\bar{y} = 120$, $s_x = 10$, $s_y = 4$, and $R = -0.6$. Compute $\hat{\beta}_1$ and $\hat{\beta}_0$ and write the fitted line.

2. A model has $R = 0.8$. What proportion of the variation in y does it explain, and what does the remaining proportion represent?

3. A two-level categorical fit is $\widehat{y} = 30 + 7\,\texttt{is\_premium}$. State the mean of each group.

4. A used car dealer collected data on $n = 32$ used cars. For each car, the dealer recorded the car's age (x, in years) and the selling price (Y, in dollars), and fitted $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The summary statistics are as follows.

$\bar{x}$	s_x	$\bar{y}$	s_y	r
6	2	14000	3000	−0.80

   (a) Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (b) A used car is $x_i = 9$ years old and its actual selling price is $y_i = 10000$ dollars. Use the model to predict the mean selling price $\hat{y}_i$ at $x_i = 9$ and compute the residual e_i.

   (c) Calculate the coefficient of determination R^2 and the residual standard error $\hat{\sigma}$.

5. A research team investigates the relationship between newborns' height (x, in cm) and weight (Y, in kg) in a city using the linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ independently. The team collects data from 16 randomly selected newborns, with summary statistics

$$\bar{x} = 49.825, \quad \bar{y} = 3.8, \quad S_{xx} = 45.45, \quad S_{yy} = 0.8096, \quad S_{xy} = 4.414.$$

   (a) Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (b) Use the regression model to predict the average weight $\hat{y}$ of a newborn whose height is $x = 50$ cm.

   (c) Define a new explanatory variable $x'_i = 10x_i$, converting height from centimeters to millimeters. Using x'_i as the predictor in $Y_i = \beta'_0 + \beta'_1 x'_i + \varepsilon'_i$, find the least squares estimates $\hat{\beta}'_0$ and $\hat{\beta}'_1$.

6. A simple linear regression model is used to explain the relationship between the total game time x (hours) and the salary y (USD 1M) of 11 players. Assume the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ independently for $i = 1, \ldots, 11$. The summary statistics are

$$\bar{x} = 20, \quad \bar{y} = 2.6727, \quad s_x = 6.4653, \quad s_y = 0.4839, \quad S_{xx} = 418, \quad S_{xy} = 26.6, \quad r = 0.8502.$$

(a) Compute the least squares estimates of the regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) Compute the fitted value $\hat{y}_4$ and residual e_4 for the $i = 4$th player, whose game time is $x_4 = 25$ and salary is $y_4 = 3.0$.

(c) Suppose the sum of squared errors is $SSE = \sum_{i=1}^{11}(y_i - \hat{y}_i)^2 = 0.6489$. Compute the mean squared error $MSE = \hat{\sigma}^2$.

7. A travel analyst collected data on $n = 30$ flights. For each flight, the analyst recorded the flight distance (x, in miles) and the flight time (y, in minutes), and fits the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The summary statistics are

$$\bar{x} = 1400, \quad s_x = 480, \quad \bar{y} = 210, \quad s_y = 60, \quad r = 0.64.$$

(a) Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) A flight has distance $x_i = 1000$ miles and actual flight time $y_i = 190$ minutes. Predict the mean flight time $\hat{y}_i$ at $x_i = 1000$ and compute the residual e_i.

(c) Calculate the coefficient of determination R^2 and the sum of squared errors $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

9.3 Inference for linear regression

A fitted slope $\hat{\beta}_1$ is computed from a sample, so it carries sampling uncertainty. The central question of regression inference is whether the true slope β_1 is really nonzero — whether x helps predict y at all — or whether the observed tilt could be noise.

9.3.1 The normality assumption

Point estimation needed only that the errors have mean 0 and constant variance. Inference needs more: a distribution for the errors.

DEFINITION 9.7 • Normal regression model

For inference we assume

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the errors are independent and identically distributed as $\varepsilon_i \sim N(0, \sigma^2)$.

9.3.2 Testing the slope

t-TEST FOR THE SLOPE β_1

Under the normal regression model, for $H_0 : \beta_1 = 0$ the statistic

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \overset{H_0}{\sim} t(n-2),$$

where $\hat{\sigma} = \sqrt{MSE}$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$. The alternative may be two- or one-sided, with p-values and rejection regions read from $t(n-2)$ exactly as in Chapter 8.

NOTE

Treat $x_1, \dots, x_n$ as fixed. Since $\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_i (x_i - \bar{x}) Y_i$ is a linear combination of the independent normal Y_i, it is itself normal, with

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} \Big[\beta_0 \sum_i (x_i - \bar{x}) + \beta_1 \sum_i (x_i - \bar{x}) x_i \Big] = \beta_1, \qquad \operatorname{Var}(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum_i (x_i - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{S_{xx}}.$$

Hence $Z := (\hat{\beta}_1 - \beta_1)/(\sigma/\sqrt{S_{xx}}) \sim N(0,1)$. It is known that $V := (n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$ independent of Z, so

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{Z}{\sqrt{V/(n-2)}} \sim t(n-2).$$

EXAMPLE 9.8 • Does queue depth predict throughput?

A reliability team has $n = 29$ days of telemetry and asks whether the average job-queue depth x predicts the day-over-day change in throughput y (percent). The hypotheses are $H_0 : \beta_1 = 0$ (queue depth is not predictive) against $H_A : \beta_1 \neq 0$. The summaries are

$$n = 29, \quad S_{xx} = 113.9, \quad S_{xy} = -101.4, \quad \hat{\sigma}^2 = 79.4.$$

The slope, its standard error, and the statistic are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-101.4}{113.9} = -0.890, \qquad SE = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{\sqrt{79.4}}{\sqrt{113.9}} = 0.835, \qquad t = \frac{-0.890}{0.835} = -1.066,$$

on $t(27)$. The two-sided p-value is $2\,\mathbb{P}(T \le -1.066) = 0.296$, far above 0.05, so we do *not* reject H_0: these data do not provide convincing evidence that queue depth predicts the throughput change (Figure 9.10).

```
> g <- lm(tput_change ~ queue_depth, data = telemetry)
> summary(g)

Call:
lm(formula = tput_change ~ queue_depth, data = telemetry)

Residuals:
 Min 1Q Median 3Q Max
-14.0124 -7.6989 0.0913 7.2974 16.1447

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.3644 5.1553 -1.429 0.165
queue_depth -0.8897 0.8350 -1.066 0.296

Residual standard error: 8.913 on 27 degrees of freedom
Multiple R-squared: 0.04035,  Adjusted R-squared: 0.004812
F-statistic: 1.135 on 1 and 27 DF, p-value: 0.2961
```

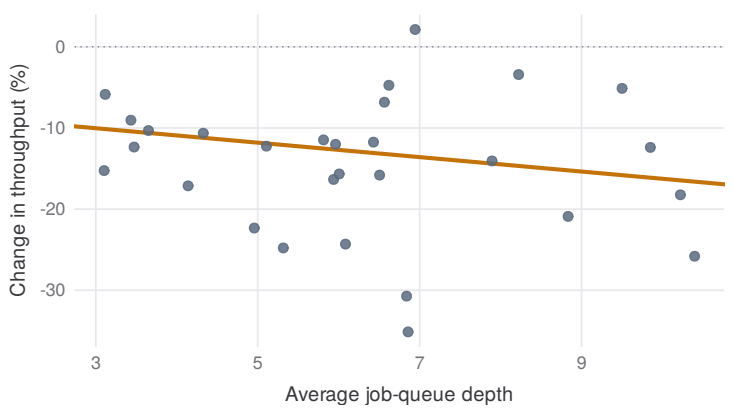


Figure 9.10. Day-over-day throughput change versus average job-queue depth over 29 days. The fitted slope is slightly negative but not statistically distinguishable from zero ($p = 0.30$).

NOTE

Regression software reports exactly the pieces of this test. In the coefficient table the `Estimate` column is $\hat{\beta}_1$, `Std. Error` is $SE(\hat{\beta}_1)$, `t value` is their ratio, and `Pr(>|t|)` is the two-sided p-value on $t(n-2)$. We rarely test the intercept, so attention stays on the predictor's row.

GUIDED PRACTICE 9.2

Use the throughput regression output from Example 9.4 to test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$ at $\alpha = 0.05$.

Answer: The `prompt_len` row gives $\hat{\beta}_1 = -0.0431$, $SE = 0.0108$, $t = -3.98$ on $t(48)$, and a two-sided p-value of 0.00023. Since $p < 0.05$, reject H_0: prompt length is a statistically significant predictor of throughput.

9.3.3 A confidence interval for the slope

t-CONFIDENCE INTERVAL FOR β_1

A $100(1-\alpha)\%$ confidence interval for the slope is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\,\frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\,SE(\hat{\beta}_1),$$

with $df = n - 2$.

EXAMPLE 9.9 • Interval for the throughput slope

With $\hat{\beta}_1 = -0.0431$, $SE = 0.0108$, and $df = 48$, the 95% critical value is $t_{0.025}(48) = 2.01$, so

$$-0.0431 \pm 2.01 \times 0.0108 = (-0.0648,\ -0.0214).$$

We are 95% confident that each additional token lowers throughput by between 0.0214 and 0.0648 requests/hour. The interval excludes 0, agreeing with the significant slope test.

```
> qt(p = 0.025, df = 48, lower.tail = FALSE)
[1] 2.010635
```

9.3.4 Conditions for the least squares line

The estimates are always computable, but the inference above is trustworthy only when four conditions hold. Each is a clause of the normal regression model, and each is checked with a residual plot.

1. **Linearity**: the mean of Y is linear in x, $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$. Leftover curvature in the residual plot signals a violation.
2. **Nearly normal residuals**: the errors ε_i are normal. Check with a histogram or normal probability plot of the residuals.
3. **Constant variability**: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for every i. A funnel-shaped residual plot signals nonconstant variance.

4. **Independent observations**: the errors are mutually independent. Beware data collected in sequence (time series), where successive errors are often correlated.

GUIDED PRACTICE 9.3

Do the queue-depth data of Example 9.8 satisfy the conditions? Inspect the residual plot in Figure 9.11.

Answer: The residuals scatter around 0 with no clear curve and roughly even spread, so linearity and constant variability look reasonable. The observations are consecutive days, so independence is the condition most worth scrutinizing. A weak fit is not the same as a violated condition: the model is adequate, the predictor simply carries little information.

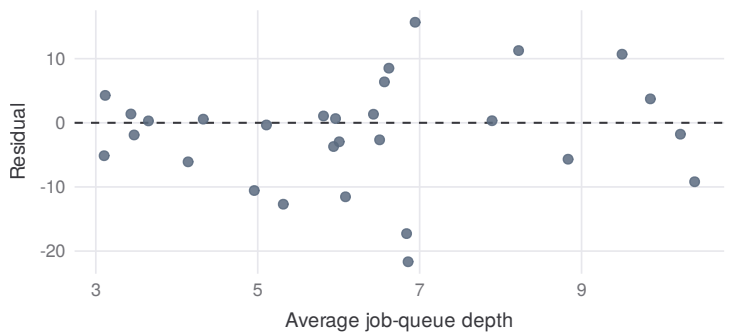


Figure 9.11. Residual plot for the queue-depth model: no obvious curve or funnel, supporting the linearity and constant-variance conditions.

GUIDED PRACTICE 9.4

A benchmarking study fits a line predicting a model's fine-tuned score y from its base score x for $n = 170$ models, with $\varepsilon_i \sim N(0, \sigma^2)$. Test $H_0 : \beta_1 = 0$ against the one-sided $H_A : \beta_1 > 0$ using

$$n = 170, \quad S_{xx} = 1157.40, \quad S_{yy} = 1010.43, \quad S_{xy} = 331.37, \quad \hat{\sigma} = 2.334.$$

Compute the slope and statistic and conclude.

Answer:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{331.37}{1157.40} = 0.286, \qquad SE = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{2.334}{\sqrt{1157.40}} = 0.0686, \qquad t = \frac{0.286}{0.0686} = 4.17,$$

on $t(168)$. The upper-tail p-value $\mathbb{P}(T \geq 4.17) \approx 2 \times 10^{-5}$ is far below 0.05, so reject H_0: a higher base score is associated with a higher fine-tuned score.

EXERCISES (§9.3)

1. A shooting instructor analyzes the relationship between a recruit's practice time (x, hours) and shooting-test score (Y, points) using the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. A

random sample of $n = 10$ recruits is analyzed using the R function lm(), with the results shown below. Answer the following.

```
> x <- c(10, 12, 14, ..., 24, 26, 28)
> y <- c(64, 80, 60, ..., 94, 80, 97)
> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
Min 1Q Median 3Q Max
-12.030 -8.614 -1.500 8.462 10.758

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.5152 11.0408 4.756 0.00143 **
x 1.3939 0.5562 2.506 0.03659 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.1 on 8 degrees of freedom
Multiple R-squared: 0.4398, Adjusted R-squared: 0.3698
F-statistic: 6.28 on 1 and 8 DF, p-value: 0.03659
```

(a) Using the R output, find the values in the table below.

Estimated regression coefficients	Estimated regression equation
$\hat{\beta}_0$	$\hat{\beta}_1$

(b) The first recruit's practice time is $x_1 = 10$ and shooting-test score is $y_1 = 64$. Using the simple linear regression analysis, find the fitted value $\hat{y}_1$ and the residual e_1.

(c) Suppose $\hat{\sigma}^2 = MSE = 102.01$. Find $\sum_{i=1}^{10} e_i^2$ for the residuals $e_1, \ldots, e_{10}$.

(d) Using the R output, we test the hypothesis $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ at significance level $\alpha = 0.05$ to see whether the practice time affects the shooting-test score.

i. The observed value of the test statistic is ( ), and its p-value is ( ). Therefore, at significance level $\alpha = 0.05$, we ( ) H_0.

ii. State the conclusion in the context of the data.

2. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

where $x_1, \cdots, x_n$ are assumed to be fixed constants. The least squares estimator of the slope β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Using the fact that

$$\sum_{i=1}^n (x_i - \overline{x}) = 0, \quad \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i - \overline{x}) x_i,$$

show that $\hat{\beta}_1$ is an unbiased estimator of β_1.

3. A research team is studying the relationship between smoking status (x, where $x = 1$ if the person is a smoker and $x = 0$ otherwise) and lung capacity (Y, measured in 1000 ml) in residents of a city, using the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ independently. A random sample of 5 smokers and 10 non-smokers was collected; the R output is shown below.

```
smoke  <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
breath <- c(30, 31, 38, 30, 35, 32, 48, 42, 35, 40, 36, 41, 47, 36, 34)
summary(lm(breath ~ smoke))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.100      1.563  25.012 2.22e-12 ***
smoke         -6.300      2.708  -2.327   0.0368 *
---
Residual standard error: 4.944 on 13 degrees of freedom
Multiple R-squared:  0.294, Adjusted R-squared:  0.2397
F-statistic: 5.414 on 1 and 13 DF,  p-value: 0.03679
```

(a) Estimate the average lung capacity for smokers ($x = 1$) and for non-smokers ($x = 0$).

(b) Compute the sample means $\bar{x} = \dfrac{1}{15}\sum_{i=1}^{15} x_i$ and $\bar{y} = \dfrac{1}{15}\sum_{i=1}^{15} y_i$. (Hint: use $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.)

(c) Compute the observed sample correlation coefficient r between x_i and y_i.

(d) We test $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 < 0$ at significance level $\alpha = 0.05$.

i. Find the null distribution of the test statistic $\dfrac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$.

ii. Write down the p-value for this test. Do you reject the null hypothesis?

9.4 Multiple linear regression

Every regression so far has used a single predictor. Yet a response rarely depends on just one thing. The time an AI system takes to answer a request depends on how long the prompt is, how long the answer is, and which accelerator runs the model; a model's accuracy depends on its size, its training data, and how it was tuned. **Multiple linear regression** extends the straight line to several predictors at once, letting each one contribute to the prediction while we hold the others fixed.

DEFINITION 9.8 • Multiple linear regression model

With p predictors $x_1, x_2, \ldots, x_p$, the **multiple linear regression model** for a numerical response Y is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

where each **coefficient** β_j measures how Y changes with x_j when the *other* predictors are held fixed, and ε is the random error. The fitted equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ is again chosen by least squares: the coefficients minimize the sum of squared residuals $\sum_i (y_i - \hat{y}_i)^2$.

The mechanics carry over almost unchanged from the single-predictor case. The one idea that is genuinely new — and the source of nearly every subtlety in this section — is the phrase *holding the*

others fixed. A coefficient in a multiple regression is no longer a stand-alone slope; it is the effect of one predictor *after the others have already done their part*.

9.4.1 Several predictors at once

EXAMPLE 9.10 • Predicting inference latency

An engineering team wants to predict the **latency** (response time, in milliseconds) of a deployed language-model API. For $n = 200$ logged requests they record the number of `output_tokens` generated, the number of `prompt_tokens` in the input, and the accelerator type (`gpu`: an older A100 or a newer H100). Fitting a multiple regression of latency on all three predictors gives:

```
> g <- lm(latency ~ output_tokens + prompt_tokens + gpu, data = serving)
> summary(g)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   150.000     14.000   10.71  < 2e-16 ***
output_tokens   7.500      0.300   25.00  < 2e-16 ***
prompt_tokens   0.120      0.030    4.00  9.1e-05 ***
gpuH100      -120.000     20.000   -6.00  6.8e-09 ***
---
Residual standard error: 45.1 on 196 degrees of freedom
Multiple R-squared: 0.883,  Adjusted R-squared: 0.881
```

The fitted equation is

$$\widehat{\texttt{latency}} = 150 + 7.5\,\texttt{output\_tokens} + 0.12\,\texttt{prompt\_tokens} - 120\,\texttt{gpuH100}.$$

To predict the latency of a request that generates 80 output tokens from a 1,000-token prompt on an H100 (so `gpuH100` = 1), substitute all three:

$$\widehat{\texttt{latency}} = 150 + 7.5(80) + 0.12(1000) - 120(1) = 750 \text{ ms}.$$

Each slope is interpreted *one predictor at a time, with the rest held constant*. Holding the prompt length and the accelerator fixed, each additional output token adds about 7.5 ms — the autoregressive cost of decoding one more token. Holding the token counts fixed, the coefficient −120 on `gpuH100` says an H100 serves an otherwise identical request about 120 ms faster than an A100. The intercept, 150 ms, is the model's prediction when every predictor is zero (an empty prompt and no output on an A100); as usual it anchors the fit rather than describing a realistic case.

INTERPRETING A COEFFICIENT "ALL ELSE HELD CONSTANT"

In a multiple regression, $\hat{\beta}_j$ is the average change in y for a one-unit increase in x_j *while every other predictor is held fixed*. It is the contribution of x_j beyond what the other predictors already explain — not the slope you would get from x_j alone.

9.4.2 Categorical predictors and indicator variables

The gpu predictor above is not a number but a category. The fit handles it exactly as in Section 9.2: it creates an **indicator variable** gpuH100 equal to 1 for an H100 and 0 for an A100. The category left at 0 — here the A100 — is the **reference level**, and every coefficient is read relative to it.

Setting the indicator to 0 or 1 splits the single fitted equation into two parallel lines, one per accelerator:

$$\begin{aligned} \text{A100 } (\texttt{gpuH100} = 0): &\quad \widehat{\texttt{latency}} = 150 + 7.5\,\texttt{output\_tokens} + 0.12\,\texttt{prompt\_tokens}, \\ \text{H100 } (\texttt{gpuH100} = 1): &\quad \widehat{\texttt{latency}} = 30 + 7.5\,\texttt{output\_tokens} + 0.12\,\texttt{prompt\_tokens}. \end{aligned}$$

The two lines share the same slopes; the indicator only shifts the intercept, lowering it by 120 ms for the H100 (Figure 9.12).

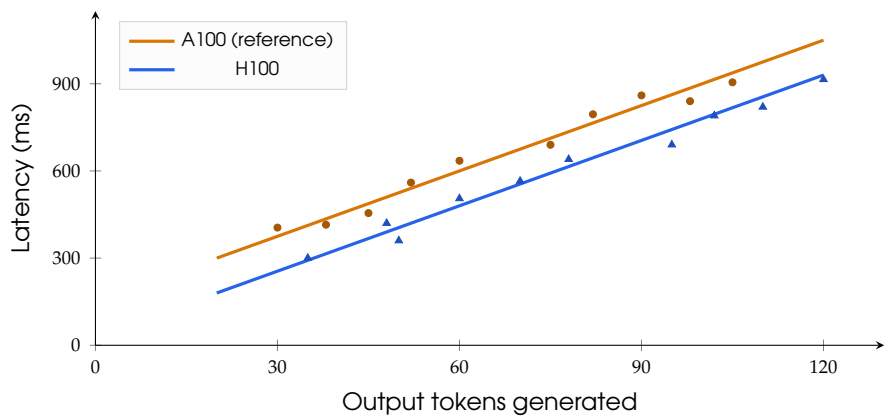


Figure 9.12. A two-level categorical predictor produces two *parallel* fitted lines. The slope on output tokens is the same for both accelerators; the indicator for the H100 only lowers the intercept — about 120 ms of latency saved at every workload.

A category with more than two levels needs more than one indicator. Three instance *families* — say CPU, GPU, and TPU — are encoded with *two* indicators (one for GPU, one for TPU), leaving CPU as the reference; in general k levels require $k - 1$ indicators. Each coefficient then compares its level against the reference, and the intercept describes the reference level.

GUIDED PRACTICE 9.5

A cost model is fit with instance family encoded as two indicators, `familyGPU` and `familyTPU`, with CPU as the reference. The estimated coefficients are +2.10 on `familyGPU` and +1.40 on `familyTPU`. Which family is cheapest, and which is most expensive, all else equal?

Answer: Both coefficients are positive, so GPU and TPU cost more than the CPU reference; CPU is cheapest. The GPU coefficient (2.10) exceeds the TPU coefficient (1.40), so GPU is the most expensive of the three.

9.4.3 Collinearity among predictors

Adding predictors can sharpen a model, but it can also muddy it. When two predictors carry nearly the same information, the fit cannot tell their effects apart, and their coefficients become unstable.

DEFINITION 9.9 • Collinearity

Two or more predictors are **collinear** when they are strongly correlated with one another. Severe collinearity inflates the standard errors of the affected coefficients and makes their estimates swing wildly with small changes in the data, because the least squares fit cannot separate effects that move together.

EXAMPLE 9.11 • Tokens and characters

Suppose the team adds `prompt_chars`, the number of characters in the prompt, to the latency model. Because text averages roughly four characters per token, `prompt_chars` is almost a constant multiple of `prompt_tokens` (Figure 9.13) — the two are about 99% correlated. The fit now has to split one real effect between two near-duplicate predictors, and neither coefficient can be pinned down: their standard errors balloon and the estimates may even flip sign, despite the model's predictions barely changing. The cure is not a clever calculation but a modeling decision: keep one of the two predictors and drop the other.

! CAUTION

Collinearity hurts *interpretation*, not necessarily *prediction*: a model with redundant predictors may still predict well, but its individual coefficients and their p-values become untrustworthy. When you care what each predictor *means*, avoid piling in variables that say the same thing.

9.4.4 Comparing models with adjusted R^2

How do we decide whether a new predictor earns its place? The obvious score, R^2, is no help: *adding any predictor can only increase* R^2, even a column of random noise, because the fit can always use

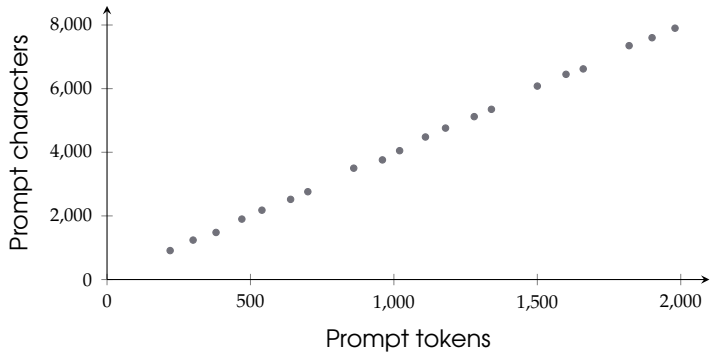


Figure 9.13. Two collinear predictors. Prompt characters track prompt tokens almost perfectly, so a model that includes both gains little and cannot disentangle their separate effects.

the extra freedom to shave the training residuals a little. A fair comparison must charge for that freedom.

DEFINITION 9.10 • Adjusted R^2

For a model with n observations and p predictors, the **adjusted** R^2 is

$$R^2_{\text{adj}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - (1-R^2)\,\frac{n-1}{n-p-1}.$$

Unlike R^2, it rises only when a new predictor improves the fit by more than would be expected by chance, and it falls when a predictor adds more complexity than signal.

EXAMPLE 9.12 • Does each predictor earn its place?

Building the latency model up one predictor at a time gives the following scores:

Model	R^2	R^2_{adj}
`output_tokens` only	0.810	0.809
+ `prompt_tokens`	0.832	0.830
+ `gpu`	0.883	0.881
+ `prompt_chars` (redundant)	0.883	0.881

Each of the first three predictors raises *both* scores: they carry real information. The redundant `prompt_chars` adds essentially nothing to R^2 (still 0.883) and does *not* raise adjusted R^2; computed to finer precision it edges down from 0.8812 to 0.8806, the penalty for the extra parameter outweighing its negligible contribution — a clear signal that it is not worth keeping.

This is the same lesson the bias–variance tradeoff will make in the next chapter, seen from the side

of model choice: more predictors are not automatically better. Among competing models, prefer the one with the higher adjusted R^2 — the **parsimonious** model that explains the most with the fewest moving parts.

CHAPTER 10

Logistic Regression

The regression of the previous chapter predicts a *numerical* response. But many of the questions an AI system must answer are not numerical at all — they are *yes-or-no*. Is this transaction fraudulent? Is this message spam? Does this scan show the condition? The target is a category, and a straight line, free to wander above one and below zero, cannot represent a probability. **Logistic regression** adapts the regression idea to a binary outcome. Section 10.1 explains why a line fails and introduces the **odds** and the **logit** that fix it, arriving at the S-shaped **logistic** curve. Section 10.2 fits the model with **R** and interprets its coefficients as **odds ratios**. Section 10.3 turns predicted probabilities into decisions and confronts the hard part — choosing a threshold — with the tools that pervade machine learning: the **confusion matrix**, **sensitivity** and **specificity**, the **ROC** curve, and cost-based **utility**. Section 10.4 then steps back to name the ideas that linear and logistic regression share with all of **supervised learning**.

10.1 Modeling a binary outcome

10.1.1 When the response is a category

A binary response is coded with the numbers 0 and 1, but those labels are names, not quantities. We are not really predicting a 0 or a 1; we are predicting the *probability* that the outcome is a 1. That shift — from a number to a probability — is what makes a line the wrong tool.

EXAMPLE 10.1 • Flagging fraudulent transactions

A payments platform logs each card transaction and later learns whether it was fraudulent. Encode the outcome as `is_fraud` = 1 for fraud and 0 otherwise, and suppose we want to predict it from the transaction `amount`. The outcomes stack at two heights, 0 and 1 (Figure 10.1): legitimate charges (mostly small) along the bottom, fraudulent ones (skewing larger) along the top. Fitting an ordinary least squares line to these 0/1 values produces the dashed line in the figure — and it is plainly wrong. At small amounts it predicts a *negative* probability of fraud; extended to large amounts it predicts a probability above 1. Neither is possible.

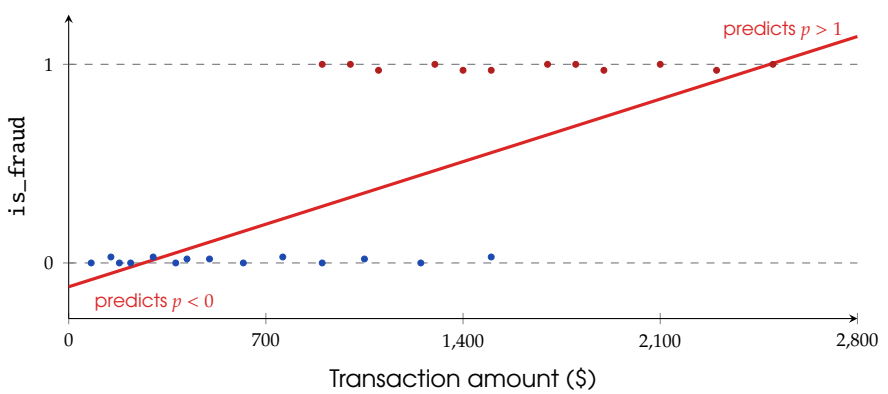


Figure 10.1. A binary response (fraud = 1, legitimate = 0) against transaction amount. The least squares line slips below 0 for small amounts and rises above 1 for large ones, so it cannot be read as a probability. We need a curve that stays inside $[0, 1]$.

10.1.2 Odds and the logit

The escape route is to model the probability on a different scale — one with no ceiling and no floor. The bridge to that scale is the **odds**.

DEFINITION 10.1 • Odds

The **odds** of an event with probability p is the ratio of the chance it happens to the chance it does not,

$$\text{odds} = \frac{p}{1-p}.$$

A probability of $p = 0.5$ is odds of 1 ("even odds"); $p = 0.8$ is odds of 4 ("4 to 1"); $p = 0.1$ is odds of $1/9$. As p ranges over $(0, 1)$, the odds range over $(0, \infty)$.

Odds remove the upper bound but keep the lower one at zero. Taking the logarithm removes that floor too, stretching the scale across the whole number line. This logarithm of the odds is the **logit**.

DEFINITION 10.2 • Logit (log-odds)

The **logit** of a probability p is the natural logarithm of its odds,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

It maps a probability in $(0, 1)$ to a number in $(-\infty, \infty)$: $p = 0.5$ gives logit 0, probabilities above 0.5 give positive logits, and probabilities below 0.5 give negative logits.

The logit converts a probability into an unbounded number, so a linear model can predict it safely. To turn a predicted logit *back* into a probability we invert the transformation, and the inverse is the function that gives logistic regression its name and its shape.

THE LOGISTIC (INVERSE-LOGIT) FUNCTION

Solving $\text{logit}(p) = z$ for p gives the **logistic function**

$$p = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^{z}}{1+e^{z}}.$$

It takes any real number z and returns a value strictly between 0 and 1 (Figure 10.2): large positive z maps near 1, large negative z near 0, and $z = 0$ maps to exactly 0.5. Its S-shape is exactly the bounded curve the scatterplot demanded.

10.1.3 The logistic regression model

We now have every piece. Put a linear model on the logit scale, and read the probability off the logistic curve.

DEFINITION 10.3 • Logistic regression model

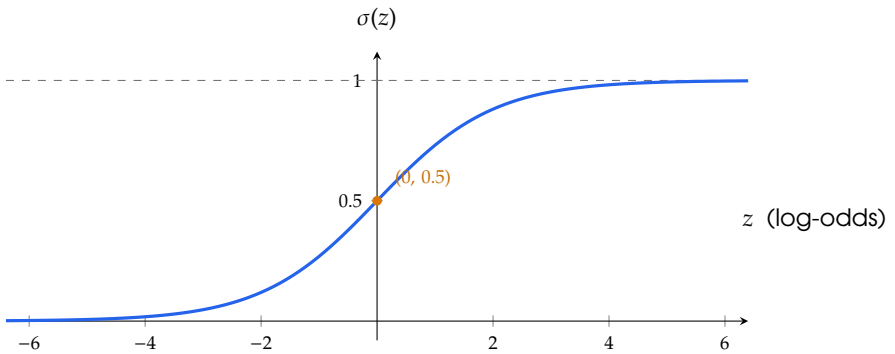


Figure 10.2. The logistic function $\sigma(z) = 1/(1 + e^{-z})$ squeezes the entire real line into the interval $(0, 1)$. A linear model predicts the log-odds z; the logistic curve converts that prediction into a probability.

Logistic regression models the probability $p = \mathbb{P}(y = 1)$ of a binary outcome through a linear predictor on the log-odds scale,

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

which is equivalent to passing the linear predictor through the logistic function,

$$p = \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) = \frac{1}{1 + \exp\big(-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\big)}.$$

REMARK 10.1 • A member of a larger family

Logistic regression is the best-known **generalized linear model** (GLM). Every GLM keeps the linear predictor $\beta_0 + \beta_1 x_1 + \cdots$ but pairs it with a response distribution and a *link function* that connects the linear predictor to that distribution's mean. Logistic regression assumes a binomial (yes/no) response and uses the logit link; swapping in a different distribution and link — a Poisson with a log link, for count data — gives a different GLM with the same machinery.

EXERCISES (§10.1)

1. For an event with probability p:
   (a) A model predicts $p = 0.8$. Find the odds and the logit of this probability.
   (b) An event has odds $1/4$. Find its probability p.
   (c) Find p when the logit equals 0, and when it equals -2 (use $\sigma(-2) \approx 0.119$).
2. A logistic model gives $\operatorname{logit}(\hat{p}) = -1.5 + 0.5\,x$.
   (a) Find the predicted probability $\hat{p}$ when $x = 5$ (use $\sigma(1.0) \approx 0.731$).
   (b) For what value of x is $\hat{p} = 0.5$?

10.2 Fitting and interpreting logistic regression

10.2.1 Fitting the model

The coefficients are estimated from data, but not by least squares. There is no closed form like the slope of Chapter 9; instead the computer searches for the coefficients that make the observed 0/1 outcomes most probable (the method of **maximum likelihood**). In **R** the only change from a linear model is to call `glm` with `family = binomial` in place of `lm`.

EXAMPLE 10.2 • A logistic fit for fraud

Fitting fraud on transaction amount for $n = 4{,}000$ logged transactions:

```
> g <- glm(is_fraud ~ amount, data = txn, family = binomial)
> summary(g)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.000000   0.420000   -9.52  < 2e-16 ***
amount       0.003000   0.000400    7.50  6.4e-14 ***
---
    Null deviance: 1184.2  on 3999  degrees of freedom
Residual deviance:  902.7  on 3998  degrees of freedom
AIC: 906.7
```

The fitted model, written on the log-odds scale, is

$$\log\!\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.00 + 0.0030\,\texttt{amount}.$$

Notice the output reports *deviance*, not R^2 or an F-statistic: those belong to least squares. Deviance plays the role of the residual sum of squares for a GLM, and smaller is better.

10.2.2 Predicted probabilities

To predict a probability, evaluate the linear part and pass it through the logistic function. For a \$1,000 transaction,

$$\hat{z} = -4.00 + 0.0030(1000) = -1.0, \qquad \hat{p} = \sigma(-1.0) = \frac{1}{1+e^{1.0}} = 0.27.$$

A \$1,500 transaction gives $\hat{z} = 0.5$ and $\hat{p} = 0.62$; a \$2,000 transaction gives $\hat{z} = 2.0$ and $\hat{p} = 0.88$. Plotting $\hat{p}$ across all amounts traces the S-curve through the data (Figure 10.3), rising smoothly from near 0 to near 1 and never leaving the interval — precisely the behavior the straight line could not manage.

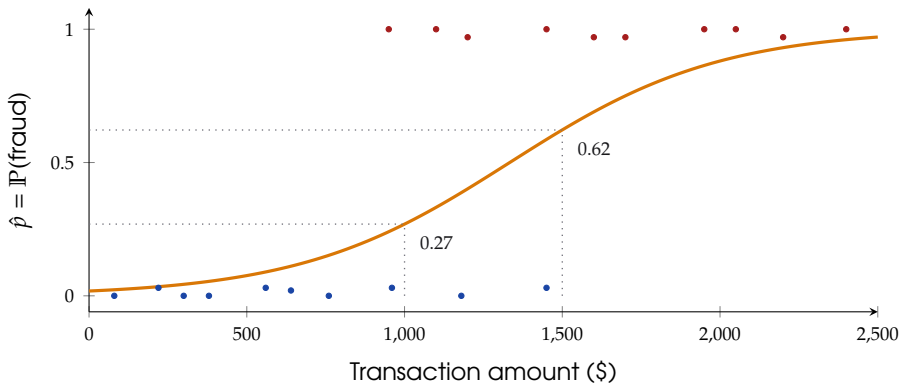


Figure 10.3. The fitted logistic model for fraud. The curve gives the predicted probability at every amount; dotted guides read off $\hat{p} = 0.27$ at \$1,000 and $\hat{p} = 0.62$ at \$1,500.

10.2.3 Odds ratios: interpreting the slope

In a linear model the slope is a change in y. Here the linear part is the log-odds, so the slope β_1 is the change in *log-odds* per unit of x — not directly meaningful. Exponentiating fixes that: e^{β_1} is the factor by which the *odds* multiply for each one-unit increase in x. This multiplier is the **odds ratio**.

> **SLOPES ARE ODDS RATIOS**
>
> In logistic regression, a one-unit increase in x_j (holding the other predictors fixed) multiplies the odds of the outcome by e^{β_j}, the **odds ratio**. A positive coefficient means $e^{\beta_j} > 1$ (the odds go up); a negative coefficient means $e^{\beta_j} < 1$ (the odds go down); a coefficient of 0 means $e^0 = 1$ (no effect).

In the fraud fit, the amount coefficient 0.0030 per dollar corresponds to 0.30 per \$100, so each additional \$100 multiplies the odds of fraud by $e^{0.30} = 1.35$ — a 35% rise in the odds. Categorical predictors enter through indicators exactly as in Section 9.4, and their odds ratios compare a level against the reference. Adding two more predictors — account age and whether the charge came from a new device — gives a richer model:

```
> g2 <- glm(is_fraud ~ amount + acct_age_yr + new_device,
+           data = txn, family = binomial)
> summary(g2)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.500000   0.550000   -4.55  5.4e-06 ***
amount         0.002800   0.000450    6.22  5.0e-10 ***
acct_age_yr   -0.450000   0.120000   -3.75  0.00018 ***
new_deviceYes  1.200000   0.300000    4.00  6.3e-05 ***
```

Reading each slope as an odds ratio, all else held constant:

- `amount`: $e^{0.0028 \times 100} = e^{0.28} = 1.32$, so each extra \$100 raises the odds of fraud by about 32%;

- `acct_age_yr`: $e^{-0.45} = 0.64$, so each additional year of account age *cuts* the odds to 64% of their value — a 36% reduction;
- `new_device`: $e^{1.20} = 3.32$, so a charge from a new device has about 3.3 times the odds of fraud of one from a known device.

Plugging an indicator into the model shifts the whole logistic curve, just as it shifted the line in Section 9.4: charges from a new device sit on a curve raised toward higher probabilities at every amount (Figure 10.4).

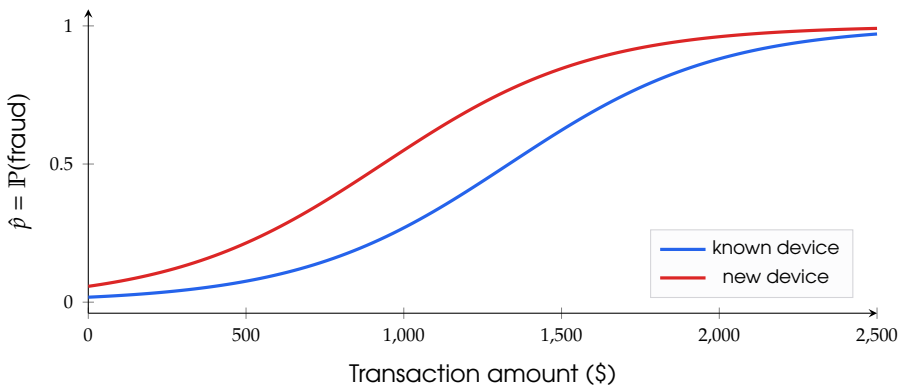


Figure 10.4. An indicator predictor shifts the entire logistic curve. At any amount, a charge from a new device carries a higher predicted probability of fraud than one from a known device.

! CAUTION

An odds ratio is *not* a ratio of probabilities. "3.3 times the odds" does not mean "3.3 times as likely." When the outcome is rare, the two are close; when it is common, they diverge sharply. Reserve the multiplicative language for *odds*, and convert to a probability (through σ) before making claims about how *likely* something is.

10.2.4 Inference for a coefficient

Each coefficient comes with a standard error, and inference mirrors the t-tests of Chapter 9 with one change: the reference distribution is the standard normal, so the test statistic is called a z rather than a t.

To test whether a predictor matters, $H_0 : \beta_j = 0$ against $H_A : \beta_j \neq 0$, form

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.$$

For the new-device indicator, $z = 1.20/0.30 = 4.00$, whose two-sided p-value ($\approx 6 \times 10^{-5}$) is far below 0.05: a new device is significantly associated with fraud.

A confidence interval is built on the log-odds scale and then exponentiated to become an interval for the *odds ratio*. A 95% interval for the new-device coefficient is

$$1.20 \pm 1.96 \times 0.30 = (0.61,\ 1.79),$$

and exponentiating each endpoint gives the odds-ratio interval

$$\left(e^{0.61},\ e^{1.79}\right) = (1.84,\ 5.98).$$

We are 95% confident that a new device multiplies the odds of fraud by somewhere between about 1.8 and 6.0 times. Because the interval lies entirely above 1, the effect is significant — the same conclusion the z-test reached.

EXERCISES (§10.2)

1. A logistic regression of loan default (1 = default) on annual income (in \$1,000s) gives

$$\text{logit}(\hat{p}) = 2.0 - 0.04\,\texttt{income}.$$

   (a) Interpret the sign of the income coefficient.
   (b) Find the odds ratio for a \$10,000 increase in income and interpret it (use $e^{-0.4} \approx 0.670$).
   (c) Predict the probability of default for an applicant with income \$50,000.

2. For the fitted fraud model $\text{logit}(\hat{p}) = -4.00 + 0.0030\,\texttt{amount}$ (amount in dollars):
   (a) Compute the predicted probability of fraud for a \$1,200 transaction (use $\sigma(-0.4) \approx 0.401$).
   (b) By what factor do the odds of fraud change for each additional \$100 in amount? (Use $e^{0.30} \approx 1.35$.)

10.3 Making and evaluating decisions

10.3.1 From probabilities to decisions

A logistic model outputs a probability, but a fraud system must ultimately *act*: flag the transaction or let it through. The simplest rule chooses a **threshold** t and flags every transaction whose predicted probability exceeds it. Comparing the flag against the truth sorts every case into one of four boxes.

DEFINITION 10.4 • Confusion matrix

Fixing a threshold turns predicted probabilities into predicted labels. Cross-tabulating predicted against actual labels gives the **confusion matrix**, with four counts: **true positives** (TP, correctly flagged), **false positives** (FP, flagged but legitimate), **true negatives** (TN, correctly passed), and **false negatives** (FN, missed fraud).

10.3.2 Sensitivity and specificity

The four counts compress into two rates that describe how well the rule catches what it should and leaves alone what it should.

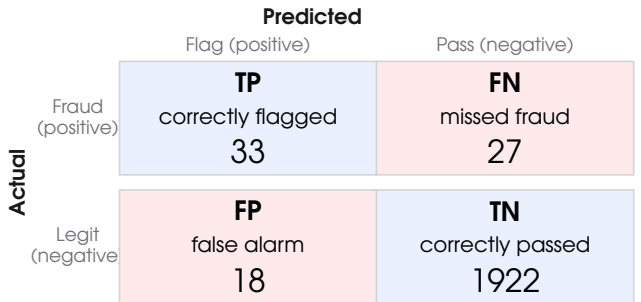


Figure 10.5. The confusion matrix at threshold $t = 0.5$ for a test set of 2,000 transactions (60 fraudulent). Teal cells are correct decisions; red cells are the two kinds of error a classifier can make.

DEFINITION 10.5 • Sensitivity and specificity

Sensitivity (the true-positive rate, or *recall*) is the fraction of actual positives the rule flags,

$$\text{Sensitivity} = \frac{TP}{TP + FN};$$

Specificity (the true-negative rate) is the fraction of actual negatives it passes,

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

At the threshold of Figure 10.5, sensitivity is $33/60 = 0.55$ (the rule catches just over half the fraud) and specificity is $1922/1940 = 0.99$ (it correctly clears almost every legitimate charge). These two move in opposite directions as the threshold changes. Lowering t flags more transactions: it catches more fraud (sensitivity up) but raises more false alarms (specificity down). Raising t does the reverse. There is no threshold that makes both perfect; every choice is a trade.

Threshold t	0.10	0.25	0.50	0.75	0.90
Sensitivity	0.92	0.78	0.55	0.33	0.18
Specificity	0.86	0.95	0.99	0.998	0.999

A useful way to see the trade is to picture the predicted-probability scores of the two groups as overlapping distributions (Figure 10.6). Legitimate charges pile up at low scores, fraud at high scores, but the two overlap. The threshold is a vertical line: everything to its right is flagged. Slide it left and the flagged region swallows more of both curves; slide it right and it releases both. The overlap is the irreducible difficulty — no line separates the curves cleanly.

10.3.3 Base rates and what a positive really means

A high-sensitivity classifier can still be wrong most of the time it raises an alarm — if the thing it looks for is rare. This is the single most misread fact about classification, and it has nothing to do with a bad model.

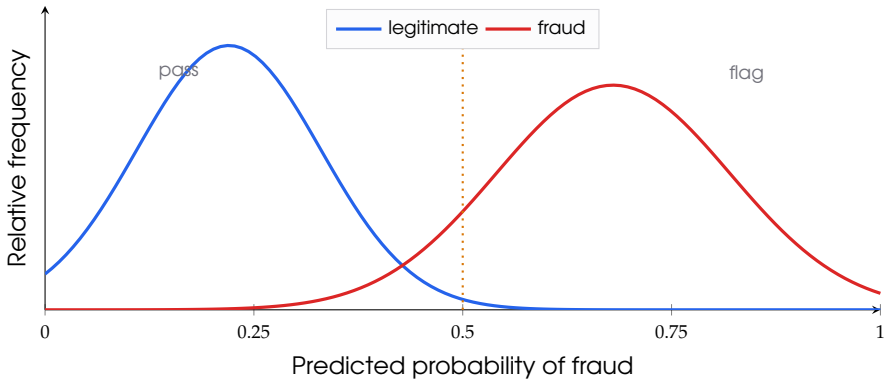


Figure 10.6. Predicted-fraud scores for the two groups overlap. The threshold (dotted) splits cases into "flag" on the right and "pass" on the left. Moving it trades sensitivity against specificity; the overlap is the part no threshold can fix.

EXAMPLE 10.3 • Why most alarms are false

Take the lenient threshold $t = 0.10$, with sensitivity 0.92 and specificity 0.86. Suppose fraud is genuinely rare: only 3% of transactions are fraudulent. Of every 10,000 transactions, about 300 are fraud and 9,700 are legitimate. The rule flags

$$0.92 \times 300 = 276 \text{ of the frauds}, \qquad (1 - 0.86) \times 9{,}700 = 1{,}358 \text{ of the legitimate charges.}$$

So it raises $276 + 1{,}358 = 1{,}634$ alarms, of which only 276 are real. The chance that a flagged transaction is *actually* fraud is

$$\mathbb{P}(\text{fraud} \mid \text{flagged}) = \frac{276}{1{,}634} \approx 0.17.$$

Despite catching 92% of fraud, fewer than one alarm in five is genuine — because the 14% false-alarm rate, applied to the huge pool of legitimate charges, swamps the true positives.

! CAUTION

Sensitivity and specificity are properties of the *test*; the chance that a positive is correct (its **precision**) also depends on the **base rate** of the positive class. When positives are rare — fraud, disease, a flagged anomaly in a stream of normal data — even an excellent classifier produces many false alarms. This is why rare-event detection is hard, and why accuracy alone is a misleading score: a model that flags *nothing* would be 97% accurate here while catching no fraud at all.

10.3.4 ROC curves and AUC

Rather than commit to one threshold, we can summarize a model across *all* of them. Sweeping the threshold from high to low and plotting the sensitivity against one minus the specificity traces the

receiver operating characteristic (ROC) curve.

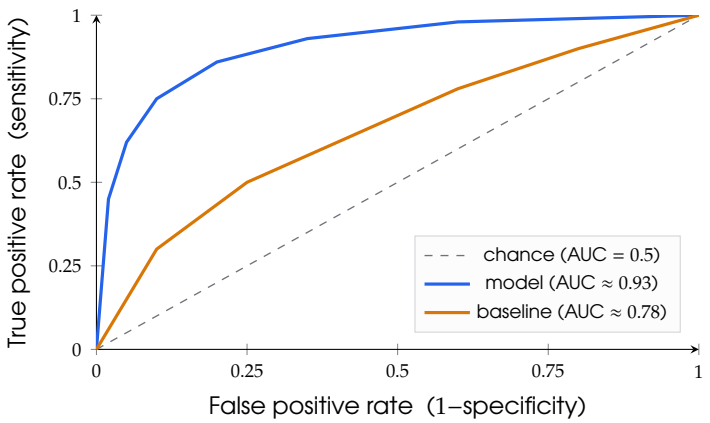


Figure 10.7. ROC curves. A perfect classifier hugs the top-left corner; the diagonal is random guessing. The stronger model bows farther toward the corner, and its larger area under the curve (AUC) confirms it dominates the baseline at every threshold.

A model that ranks fraud above legitimate charges has an ROC curve that bows toward the top-left corner; random guessing lies on the diagonal. The single number that summarizes the curve is the **area under the curve** (AUC).

> **AREA UNDER THE ROC CURVE (AUC)**
>
> The **AUC** is the area under the ROC curve. It ranges from 0 to 1: a value of 0.5 is chance-level ranking (the diagonal), values above 0.5 indicate useful ranking up to 1.0 for a perfect ranker, and values below 0.5 indicate systematically *reversed* ranking (flipping the scores would beat chance). The AUC equals the probability that the model assigns a higher score to a randomly chosen positive than to a randomly chosen negative, counting a tie as half. Because it averages over all thresholds, AUC compares models without committing to one decision rule.

In Figure 10.7 the richer model (AUC ≈ 0.93) sits above the baseline (AUC ≈ 0.78) everywhere, so it is the better *ranker* regardless of the threshold finally chosen.

10.3.5 Choosing a threshold with costs

AUC says which model ranks better, but a deployed system still needs one threshold — and the right one depends on what each outcome is *worth*. A missed fraud costs a chargeback; a false alarm costs an annoyed customer and a support call. Assigning each outcome a value turns threshold selection into an optimization.

> **EXAMPLE 10.4 • A cost-weighted threshold**

Suppose catching fraud is worth +20, correctly clearing a charge +1, a false alarm −5, and a missed fraud −100 (the chargeback dominates). The **utility** of a threshold is

$$U(t) = 20\,TP(t) + 1\,TN(t) - 5\,FP(t) - 100\,FN(t),$$

each count read from the confusion matrix at that threshold. Evaluating U across thresholds traces a curve (Figure 10.8) that peaks well below 0.5: because a missed fraud is so expensive, the best rule is *lenient*, accepting many false alarms to avoid letting fraud through. The threshold that maximizes total utility — near $t \approx 0.2$ here — is the one to deploy, not the default 0.5.

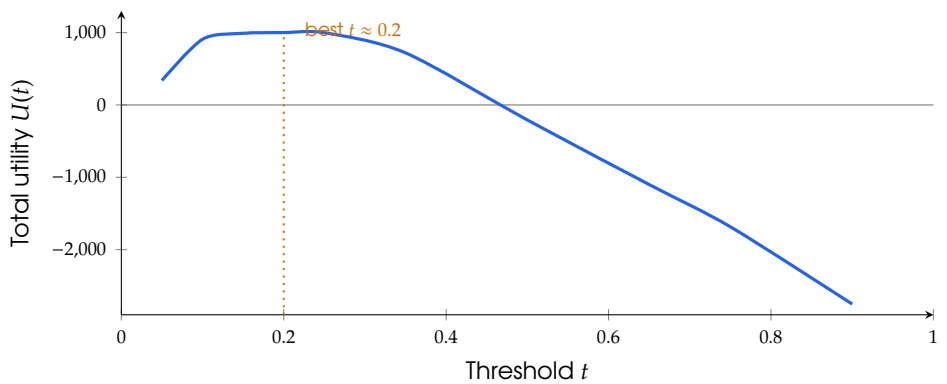


Figure 10.8. Total utility against threshold for the cost structure of Example 10.4. The peak sits near $t \approx 0.2$, far below the naive 0.5: when missing fraud is far costlier than a false alarm, the optimal rule errs toward flagging.

REMARK 10.2 • The threshold is a business decision

Nothing in the data fixes the threshold; the costs do. A spam filter, a medical screen, and a fraud detector might share the same logistic machinery yet sit at wildly different thresholds, because the price of a false positive relative to a false negative differs in each. Fitting the model is statistics; choosing where to cut is judgment informed by consequences.

EXERCISES (§10.3)

1. A classifier is applied to 500 test cases (100 actual positives, 400 actual negatives) at a fixed threshold. The confusion matrix has $TP = 80$, $FN = 20$, $FP = 40$, and $TN = 360$.
   (a) Compute the sensitivity and the specificity.
   (b) Compute the accuracy.
   (c) Compute the precision (the fraction of flagged cases that are truly positive).
2. A medical test has sensitivity 0.95 and specificity 0.90, and the disease affects 2% of the population.
   (a) Among 10,000 people, how many true positives and how many false positives does the test produce?
   (b) What is the probability that a person who tests positive actually has the disease? Comment on the result.

3. Two models are compared by their ROC curves: model A has AUC 0.80 and model B has AUC 0.55.
   (a) Which model ranks a random positive above a random negative more reliably?
   (b) What AUC corresponds to random guessing, and what would an AUC below 0.5 indicate?

10.4 From regression to machine learning

Logistic regression gave us our first *classifier*: a model that turns predictors into a decision about a category. Linear regression, two chapters running, did the same for a numerical target. Both are instances of one idea that organizes nearly all of modern prediction — *supervised learning*. Naming it makes clear which parts of these chapters carry over unchanged to far more elaborate models, and which two ideas, invisible in a single fitted curve but central to all of machine learning, are genuinely new.

10.4.1 Supervised learning

DEFINITION 10.6 • Supervised learning

In **supervised learning** we are given n labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$ and seek a function f mapping the **features** (the predictors x) to the **label** (the target y), chosen so that $f(x)$ predicts the label of a new, unseen case. When the label is numerical the task is **regression**; when it is categorical it is **classification**.

Linear regression is the prototype. The single feature is the predictor x, the label is the response y, and the learned function is the fitted line

$$f(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Least squares (Section 9.2) is the *learning* step — it chooses f from the data — and substituting a new x is the *prediction* step. The classifier that labels an image correct or incorrect (Chapter 2) is instead a classification problem, its label categorical rather than numerical. The word *supervised* refers to the labels: every training example arrives with the answer y_i attached. **Unsupervised learning**, by contrast, has no labels and searches for structure in the features alone, as when similar cases are grouped into clusters.

10.4.2 Training error, test error, and generalization

A fitted line can be judged two ways: by how well it matches the data used to fit it, and by how well it predicts data it has never seen. Only the second truly matters, because the purpose of a model is to predict new cases, not to memorize old ones.

DEFINITION 10.7 • Training and test sets

The **training set** is the data used to fit (estimate) the model. A separate **test set**, not used in fitting, is held out to measure how well the model predicts new cases. Partitioning the available data into these two parts is a **train–test split**.

This is not a new device in disguise: the held-out benchmark of Example 9.1 is exactly a test set. A model's score on the *public* benchmark — whose questions circulate openly, and may have leaked into training — can be optimistically inflated, whereas its score on the *private* benchmark estimates how it will perform in the wild. Measuring the average squared residual on each set gives a *training error* and a *test error*; because the coefficients were chosen to make the training residuals small, the training error is systematically the more flattering of the two.

EVALUATE ON HELD-OUT DATA

A model's error on its own training data underestimates its error on new data, because the fit was tuned to that data. The honest measure of predictive quality is the **generalization** error — the error on a test set the model never saw while being fit.

10.4.3 Overfitting and underfitting

With only a slope and an intercept, a line has very little freedom, so it can rarely chase noise. The opposite danger is a model so flexible that it bends to match every wiggle in the training data.

DEFINITION 10.8 • Overfitting and underfitting

A model **overfits** when it is flexible enough to fit the noise in the training data, achieving small training error but large test error, because the noise it absorbed does not recur in new data. A model **underfits** when it is too rigid to capture the real pattern, giving large error on both the training and the test set.

Both failures are easy to picture. Fitting a straight line to the saturating relationship of Figure 9.2 (Chapter 9) *underfits*: the line is too rigid for a curved trend and is systematically wrong. At the other extreme, imagine fitting a wiggly high-degree polynomial that threads through all n points exactly. Its training error is zero, yet it swings wildly between the points and predicts new cases poorly — a textbook *overfit*.

! CAUTION

Small training error is not a goal in itself. A curve forced through every training point typically predicts new points *worse* than a simpler line. Judge a model by its held-out performance, never by how closely it reproduces the data it was fit on.

10.4.4 The bias–variance tradeoff

Why not simply keep adding flexibility until the training error vanishes? Because the expected error on a new case splits into two competing pieces, and flexibility pushes them in opposite directions.

DEFINITION 10.9 • Bias and variance of a prediction

Imagine predicting the response at a fixed input x_0 with an estimate $\hat{f}(x_0)$ built from a random training set.

- The **bias** is the gap between the average prediction (over many training sets) and the truth, $\mathbb{E}[\hat{f}(x_0)] - f(x_0)$. It is the systematic error of a model too simple to capture the pattern.
- The **variance** is $\mathrm{Var}(\hat{f}(x_0))$: how much the prediction jumps from one training set to another. It measures sensitivity to the particular sample drawn.

BIAS–VARIANCE DECOMPOSITION

For squared-error prediction at a point x_0, the expected test error splits into three nonnegative parts,

$$\underbrace{\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right]}_{\text{expected test error}} = \underbrace{\left(\mathrm{Bias}[\hat{f}(x_0)]\right)^2}_{\text{too simple}} + \underbrace{\mathrm{Var}[\hat{f}(x_0)]}_{\text{too sensitive}} + \underbrace{\sigma^2}_{\text{irreducible noise}},$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$ is the noise inherent in y and cannot be removed by any model.

Increasing flexibility lowers bias — the model can follow the true pattern — but raises variance, because it also follows the sample's noise; reducing flexibility does the reverse. Underfitting is the high-bias extreme, overfitting the high-variance extreme. The best model minimizes the *sum*, not either term alone, so the test error typically falls and then rises as flexibility grows, tracing the U-shaped curve of Figure 10.9 while the training error keeps dropping toward zero.

REMARK 10.3 • The same logic as estimation

This is the prediction-side echo of a theme from Chapter 7: the quality of an estimator combines its bias and its variance. A more complex model trades one for the other, and good modeling — like good estimation — balances the two rather than driving just one of them to zero.

GUIDED PRACTICE 10.1

A team reports that a highly flexible model achieves almost zero error on its training images but markedly higher error on a held-out test set. Name the phenomenon, say whether bias or variance dominates, and suggest one remedy.

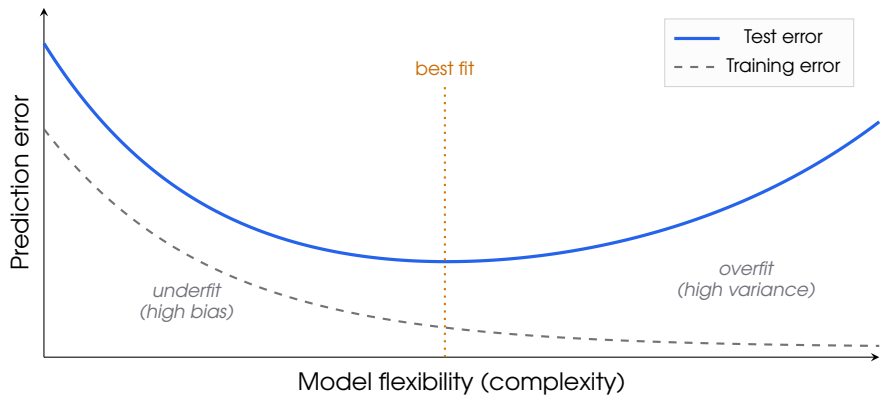


Figure 10.9. The bias–variance tradeoff. As model flexibility grows, training error falls steadily, but test error first falls (as bias shrinks) and then rises (as variance grows). The best model sits at the bottom of the U, balancing the two — not at the right edge, where training error is smallest.

Answer: This is *overfitting*: near-zero training error with high test error is the signature of high *variance*. Remedies include using a simpler (less flexible) model or collecting more training data, both of which reduce variance.

EXERCISES (§10.4)

1. Explain why a model's accuracy measured on its own training data tends to overstate how accurately it will classify new, unseen data.
2. A flexible model achieves 99% accuracy on the training set but 72% on a held-out test set. A simpler model achieves 85% on the training set and 83% on the test set.
   (a) Which model is overfitting, and how can you tell?
   (b) Which model would you expect to generalize better to new data, and why?
3. In one or two sentences, describe the bias–variance tradeoff. State whether increasing model flexibility tends to increase or decrease the bias, and whether it tends to increase or decrease the variance.

APPENDIX A

R Basics

Every numerical result in this book is accompanied by a short block of **R** output. This appendix gives the minimum needed to *reproduce* those computations yourself, rather than only read them. **R** is a free, open-source language for statistical computing (https://www.r-project.org); nothing below assumes prior programming experience.

A.1 Installing and running R

Download and install **R** for your operating system from the Comprehensive R Archive Network, https://cran.r-project.org. Most users also install the free **RStudio** desktop application (https://posit.co/download), which wraps an editor, a plots pane, and a help browser around the same **R** engine.

When **R** starts it shows a **console** with a `>` prompt: you type an expression, press Enter, and **R** prints the result. Text after a `#` is a **comment** and is ignored. Typing `?mean` (or `help(mean)`) opens the documentation for any function.

```
> 3 + 4 * 2        # arithmetic; * and / bind before + and -
[1] 11
> sqrt(16)         # a function is called as name(arguments)
[1] 4
```

The `[1]` that precedes output is just the index of the first value on the line; it is not part of the answer.

A.2 Values, vectors, and assignment

Use `<-` to store a value in a named variable, and `c()` ("combine") to build a **vector**, the basic data structure of **R**. Arithmetic and most functions act on a whole vector at once.

```
> x <- c(2, 4, 6, 8, 10)   # store a vector in x
> x
[1]  2  4  6  8 10
> mean(x)
[1] 6
> sd(x)                    # sample standard deviation
[1] 3.162278
> x * 2                    # operations apply element by element
[1]  4  8 12 16 20
> x[3]                     # the third element (R counts from 1)
[1] 6
> sum(x); length(x)        # ; separates two commands on one line
[1] 30
[1] 5
```

Two shortcuts build regular sequences: `a:b` counts by ones, and `seq()` takes an explicit step.

```
> 1:5
[1] 1 2 3 4 5
> seq(0, 1, by = 0.25)
[1] 0.00 0.25 0.50 0.75 1.00
```

A.3 Data frames

A **data frame** is the rectangular "data matrix" of Chapter 1: one row per case, one named column per variable. Build one with `data.frame()` or read a spreadsheet with `read.csv("file.csv")`. Extract a column with the `$` operator, and preview the top rows with `head()`.

```
> d <- data.frame(hours = c(1, 2, 3, 4, 5),
+                 score = c(2, 4, 5, 4, 5))
> head(d, 3)          # first 3 rows
  hours score
1     1     2
2     2     4
3     3     5
> d$score             # the score column as a vector
[1] 2 4 5 4 5
```

A statement that runs past one line continues at the `+` prompt, as in the `data.frame()` call above.

A.4 Statistical functions used in this book

Each named distribution comes with a family of four functions sharing one root: `d` for the density or pmf, `p` for the (lower-tail) cumulative probability $\mathbb{P}(X \le x)$, `q` for the quantile (the inverse of `p`), and `r` for random draws.

Distribution	density / pmf	CDF $\mathbb{P}(X \le x)$	quantile	random
Normal	`dnorm`	`pnorm`	`qnorm`	`rnorm`
Student's t	`dt`	`pt`	`qt`	`rt`
Binomial	`dbinom`	`pbinom`	`qbinom`	`rbinom`
F	`df`	`pf`	`qf`	`rf`
Chi-squared	`dchisq`	`pchisq`	`qchisq`	`rchisq`

These reproduce the probabilities and critical values used in Chapters 5–8. Pass `lower.tail = FALSE` for an upper-tail probability $\mathbb{P}(X > x)$.

```
> pnorm(1.96)                      # \Prob(Z <= 1.96), standard normal
[1] 0.9750021
> qnorm(0.975)                     # 97.5th percentile of N(0, 1)
[1] 1.959964
> qt(0.975, df = 8)                # t critical value, 8 d.f.
[1] 2.306004
> dbinom(8, size = 10, prob = 0.6)  # \Prob(X = 8) for X ~ B(10, 0.6)
[1] 0.1209324
```

The procedures of later chapters are single function calls. The most common are collected below.

Task	R function
One-/two-sample t test and CI for a mean	`t.test`
Test and CI for a proportion	`prop.test`
Sample correlation and its test	`cor`, `cor.test`
Fit a linear regression	`lm`
Coefficient table, R^2, F test	`summary`
Confidence intervals for coefficients	`confint`
Numerical summaries	`mean`, `sd`, `var`, `summary`

A.5 Reproducing a regression

The output blocks in Chapter 9 all come from `lm()` followed by `summary()`. Here is a complete, runnable example on a tiny data set.

```
> hours <- c(1, 2, 3, 4, 5)
> score <- c(2, 4, 5, 4, 5)
> fit <- lm(score ~ hours)      # fit score = b0 + b1 * hours
> summary(fit)

Call:
lm(formula = score ~ hours)
```

```
Residuals:
   1    2    3    4    5
-0.8  0.6  1.0 -0.6 -0.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2000     0.9381   2.345    0.101
hours         0.6000     0.2828   2.121    0.124

Residual standard error: 0.8944 on 3 degrees of freedom
Multiple R-squared:    0.6, Adjusted R-squared: 0.4667
F-statistic:    4.5 on 1 and 3 DF,  p-value: 0.124
```

Reading the output with the notation of Section 9.3: the `Estimate` column gives $\hat{\beta}_0 = 2.2$ and $\hat{\beta}_1 = 0.6$, so the fitted line is $\hat{y} = 2.2 + 0.6\,x$; `Std. Error`, `t value`, and `Pr(>|t|)` are $SE(\hat{\beta}_1)$, the t statistic, and its two-sided p-value; `Residual standard error` is $\hat{\sigma} = \sqrt{MSE}$; and `Multiple R-squared` is the R^2 of Definition 9.6. A confidence interval for the slope (the quantity of Example 9.9) is one more call:

```
> confint(fit, "hours", level = 0.95)
                2.5 %   97.5 %
hours      -0.3001309 1.500131
```

APPENDIX B

Statistical Tables

These reference tables give commonly used critical values and cumulative probabilities for an introductory statistics textbook.

Notation. For the standard normal table, $\Phi(z) = \mathbb{P}(Z \le z)$. For the t, chi-square, and F tables, upper-tail critical values are reported:

$$\mathbb{P}(T_\nu > t_\alpha(\nu)) = \alpha, \qquad \mathbb{P}(\chi^2_\nu > \chi^2_\alpha(\nu)) = \alpha, \qquad \mathbb{P}(F_{i,k} > F_\alpha(i,k)) = \alpha.$$

In the F tables, i is the numerator degrees of freedom and k is the denominator degrees of freedom.

Common z values for confidence intervals

Two-sided critical values $z_{\alpha/2}$ for a confidence level $1-\alpha$, where $\Phi(z_{\alpha/2}) = 1-\alpha/2$. A $100(1-\alpha)\%$ confidence interval for a mean uses $\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$.

Confidence level $1-\alpha$	80%	90%	95%	98%	99%	99.9%
$z_{\alpha/2}$	1.282	1.645	1.960	2.326	2.576	3.291

Standard normal distribution

Cumulative probabilities $\Phi(z) = \mathbb{P}(Z \leq z)$. *For negative values, use* $\Phi(-z) = 1 - \Phi(z)$.

Table B.1. Standard normal cumulative probabilities, $\Phi(z) = \mathbb{P}(Z \leq z)$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Student's t distribution

Upper-tail critical values $t_\alpha(\nu)$, where $\mathbb{P}(T_\nu > t_\alpha(\nu)) = \alpha$.

Table B.2. Upper-tail critical values of Student's t distribution, $t_\alpha(\nu)$.

ν	0.25	0.10	0.05	0.025	0.01	0.00833	0.00625	0.005
1	1.000	3.078	6.314	12.706	31.821	38.204	50.923	63.657
2	0.816	1.886	2.920	4.303	6.965	7.650	8.860	9.925
3	0.765	1.638	2.353	3.182	4.541	4.857	5.392	5.841
4	0.741	1.533	2.132	2.776	3.747	3.961	4.315	4.604
5	0.727	1.476	2.015	2.571	3.365	3.534	3.810	4.032
6	0.718	1.440	1.943	2.447	3.143	3.288	3.521	3.707
7	0.711	1.415	1.895	2.365	2.998	3.128	3.335	3.499
8	0.706	1.397	1.860	2.306	2.896	3.016	3.206	3.355
9	0.703	1.383	1.833	2.262	2.821	2.934	3.111	3.250
10	0.700	1.372	1.812	2.228	2.764	2.870	3.038	3.169
11	0.697	1.363	1.796	2.201	2.718	2.820	2.981	3.106
12	0.695	1.356	1.782	2.179	2.681	2.780	2.934	3.055
13	0.694	1.350	1.771	2.160	2.650	2.746	2.896	3.012
14	0.692	1.345	1.761	2.145	2.624	2.718	2.864	2.977
15	0.691	1.341	1.753	2.131	2.602	2.694	2.837	2.947
16	0.690	1.337	1.746	2.120	2.583	2.673	2.813	2.921
17	0.689	1.333	1.740	2.110	2.567	2.655	2.793	2.898
18	0.688	1.330	1.734	2.101	2.552	2.639	2.775	2.878
19	0.688	1.328	1.729	2.093	2.539	2.625	2.759	2.861
20	0.687	1.325	1.725	2.086	2.528	2.613	2.744	2.845
21	0.686	1.323	1.721	2.080	2.518	2.602	2.732	2.831
22	0.686	1.321	1.717	2.074	2.508	2.591	2.720	2.819
23	0.685	1.319	1.714	2.069	2.500	2.582	2.710	2.807
24	0.685	1.318	1.711	2.064	2.492	2.574	2.700	2.797
25	0.684	1.316	1.708	2.060	2.485	2.566	2.692	2.787
26	0.684	1.315	1.706	2.056	2.479	2.559	2.684	2.779
27	0.684	1.314	1.703	2.052	2.473	2.553	2.676	2.771
28	0.683	1.313	1.701	2.048	2.467	2.547	2.669	2.763
29	0.683	1.311	1.699	2.045	2.462	2.541	2.663	2.756
30	0.683	1.310	1.697	2.042	2.457	2.536	2.657	2.750
40	0.681	1.303	1.684	2.021	2.423	2.499	2.616	2.704
60	0.679	1.296	1.671	2.000	2.390	2.463	2.575	2.660
120	0.677	1.289	1.658	1.980	2.358	2.428	2.536	2.617
∞	0.675	1.282	1.645	1.960	2.327	2.394	2.498	2.576

Chi-square distribution

Upper-tail critical values $\chi^2_\alpha(\nu)$*, where* $\mathbb{P}(\chi^2_\nu > \chi^2_\alpha(\nu)) = \alpha$.

Table B.3. Upper-tail critical values of the chi-square distribution, $\chi^2_\alpha(\nu)$.

ν	0.99	0.975	0.95	0.90	0.50	0.10	0.05	0.025	0.01
1	0.0002	0.001	0.004	0.02	0.45	2.71	3.84	5.02	6.63
2	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21
3	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34
4	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28
5	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09
6	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81
7	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48
8	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09
9	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67
10	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21
11	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72
12	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22
13	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69
14	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14
15	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58
16	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00
17	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41
18	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81
19	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19
20	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57
21	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93
22	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29
23	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64
24	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98
25	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31
26	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64
27	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96
28	13.56	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28
29	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59
30	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89
40	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69
50	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15
60	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38
70	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.43
80	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33
90	61.75	65.65	69.13	73.29	89.33	107.57	113.15	118.14	124.12
100	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81

F distribution

Upper-tail critical values $F_\alpha(i,k)$, where $\mathbb{P}(F_{i,k} > F_\alpha(i,k)) = \alpha$. Columns give numerator degrees of freedom i; rows give denominator degrees of freedom k.

Upper-tail critical values for $\alpha = 0.10$

Table B.4. F distribution upper-tail critical values, $\alpha = 0.10$, panel A: numerator degrees of freedom $i = 1, \ldots, 10$.

$k\backslash i$	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65

Upper-tail critical values for $\alpha = 0.05$

Table B.5. F distribution upper-tail critical values, $\alpha = 0.05$, panel A: numerator degrees of freedom $i = 1, \ldots, 10$.

$k \backslash i$	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Upper-tail critical values for $\alpha = 0.025$

Table B.6. F distribution upper-tail critical values, $\alpha = 0.025$, panel A: numerator degrees of freedom $i = 1, \ldots, 10$.

$k \backslash i$	1	2	3	4	5	6	7	8	9	10
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05

APPENDIX C

Solutions to Exercises

Worked solutions to the exercises. Within each section, the solutions are listed in the same order as the exercises.

Chapter 1

§1.1 Statistics and artificial intelligence

1. *Descriptive* statistics summarize and describe the data actually observed, whereas *inferential* statistics use a sample to draw conclusions about a larger population. For one class's final-exam scores: a descriptive summary is, for example, the class average (or a histogram of the scores); an inferential claim is using those scores to estimate the average for a broader population, e.g. "students taking this course in general would average about this score."

2. **(1)** The population is all cadets in the regiment; the sample is the 50 cadets who are timed. **(2)** Descriptive — it merely summarizes the 50 measured times. **(3)** Inferential — it generalizes from the sample to the regiment-wide average.

§1.2 Data and variables

1. **(1)** 120 rows and 3 columns. **(2)** A single row represents one server (its three recorded attributes). **(3)** Region is categorical (nominal); number of CPU cores is numerical (discrete); average daily uptime (percent) is numerical (continuous).

2. **(1)** Numerical, continuous (it can take any value in a range). **(2)** Numerical, discrete (a count). **(3)** Categorical, ordinal (the ranks have a natural order). **(4)** Categorical, nominal (a label; arithmetic on serial numbers is meaningless). **(5)** Numerical, continuous.

§1.3 Populations, samples, and study design

1. **(1)** $\mu = 14.2$ is the parameter (it describes the whole population); $\bar{x} = 13.8$ is the statistic (computed from the sample). **(2)** Measuring the entire population is usually too costly, time-consuming, or impractical, so we study a sample. **(3)** A sample may fail to represent its population if it is selected in a biased (non-random) way — for example, a convenience or volunteer sample, or one that is too small.

2. **(1)** Observational; no causal conclusion is warranted, because patients were not randomly assigned and confounding factors may explain the shorter stays. **(2)** Experimental; a causal conclusion is warranted, because randomization balances confounders between the vaccine and placebo groups. **(3)** Observational; no causal conclusion is warranted (a lurking variable such as city size could drive both).

Chapter 2

§ 2.1 Examining numerical data

1. Draw the box from $Q_1 = 206.75$ to $Q_3 = 270$ with the median line at 248. Here IQR $= 270 - 206.75 = 63.25$, so the fences are $Q_1 - 1.5\,\text{IQR} \approx 111.9$ and $Q_3 + 1.5\,\text{IQR} \approx 364.9$. The minimum (158) and maximum (329) both lie inside the fences, so there are no outliers and the whiskers extend to 158 and 329.

2. **(1)** $Q_1 = 50$. **(2)** IQR $= Q_3 - Q_1 = 78 - 50 = 28$. **(3)** 25 soldiers (those at or above $Q_3 = 78$). **(4)** 22 sit-ups (the minimum). **(5)** 50 soldiers scored 61 or fewer.

3. **(1) O.** The median line is drawn at 55 days. **(2) O.** IQR $= Q_3 - Q_1 = 57 - 54 = 3$. **(3) X.** The smallest non-outlier (the left whisker) is at 52.5; the value 54 is Q_1, not the minimum. **(4) X.** Since 55 is the median, about 50% (not 25%) of the tires lie below it. **(5) X.** The single high outlier (62.5) and the longer right whisker indicate a *right*-skewed distribution.

4. **(1)** The median line is at 88. **(2)** IQR $= Q_3 - Q_1 = 184 - 74 = 110$. **(3)** $Q_3 + 1.5\,\text{IQR} = 184 + 1.5(110) = 349$. **(4)** The two points beyond the upper whisker (355 and 778) are outliers, so there are 2. **(5)** $778 - 63 = 715$.

5. **(1)** Draw the box from $Q_1 = 30$ to $Q_3 = 80$ with the median line at 40. Since there are no outliers, the whiskers extend to the minimum (10) and the maximum (120). **(2) (B).** The mean (50) exceeds the median (40), and the right portion (Q_3 to maximum, $80 \to 120$) is longer than the left (minimum to Q_1, $10 \to 30$); both indicate a *right*-skewed distribution.

§ 2.2 Considering categorical data

1. Recall (dog) $= \dfrac{210}{250} = 0.84$: of all images that are truly dogs, 84% are correctly predicted as dogs. Precision (dog) $= \dfrac{210}{236} \approx 0.8898$: of all images predicted to be dogs, about 88.98% are truly dogs.

2. $\dfrac{127}{600} \approx 0.2117$. This is a *joint* proportion (the proportion of images that are simultaneously truly a bird and predicted to be a bird).

Chapter 3

§ 3.1 Defining probability

1. (a) Even $= \{2,4,6\}$, so $\mathbb{P} = \dfrac{3}{6} = \dfrac{1}{2}$. (b) Greater than $4 = \{5,6\}$, so $\mathbb{P} = \dfrac{2}{6} = \dfrac{1}{3}$. (c) Even *or* greater than $4 = \{2,4,5,6\}$, so $\mathbb{P} = \dfrac{4}{6} = \dfrac{2}{3}$. (Equivalently, $\frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}$, subtracting $\mathbb{P}(\{6\}) = \frac{1}{6}$.)

2. Write $A \cup B = A \cup (B \cap A^c)$, a union of disjoint sets, so by additivity $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Also $B = (A \cap B) \cup (B \cap A^c)$ is a disjoint union, so $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c)$, giving $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Substituting yields $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

3. Each singleton has probability $\frac{1}{4}$, and $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$. The pairwise intersections are $A \cap B = A \cap C = B \cap C = \{1\}$, each with probability $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$, so A, B, C are pairwise independent. However $A \cap B \cap C = \{1\}$ has probability $\frac{1}{4}$, whereas $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{8}$; since $\frac{1}{4} \neq \frac{1}{8}$, the three events are *not* mutually independent.

4. By independence, $\mathbb{P}(\text{none fail}) = (0.97)^5 \approx 0.8587$, so

$$\mathbb{P}(\text{at least one fails}) = 1 - (0.97)^5 \approx 1 - 0.8587 = 0.1413.$$

5. Apply the two-event addition rule twice. First, $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}\big((A \cup B) \cap C\big)$. Since $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ and $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ gives $\mathbb{P}\big((A \cup B) \cap C\big) = \mathbb{P}(A \cap C) + \mathbb{P}(B \cap C) - \mathbb{P}(A \cap B \cap C)$, substituting yields

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

§ 3.2 Conditional probability

1. **(1)** Let A be the event that a camera's sensor is defective, and B be the event that the camera loses its picture. Then

$$\mathbb{P}(A) = 0.03, \quad \mathbb{P}(A^c) = 0.97, \quad \mathbb{P}(B \mid A) = 0.8, \quad \mathbb{P}(B \mid A^c) = 0.04.$$

By the law of total probability,

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(A)\,\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\,\mathbb{P}(B \mid A^c) \\ &= 0.03 \times 0.8 + 0.97 \times 0.04 = 0.024 + 0.0388 = 0.0628.\end{aligned}$$

(2) Since A and A^c form a partition of the sample space, by Bayes' theorem,

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(A)\,\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\,\mathbb{P}(B \mid A^c)} \\ &= \frac{0.03 \times 0.8}{0.03 \times 0.8 + 0.97 \times 0.04} = \frac{0.024}{0.0628} \approx 0.3822.\end{aligned}$$

2. **(1)** Let A be the event that the enemy launches an attack, and B be the event that the radar system detects an alarm. Then

$$\mathbb{P}(A) = 0.01, \quad \mathbb{P}(A^c) = 0.99, \quad \mathbb{P}(B \mid A) = 0.98, \quad \mathbb{P}(B \mid A^c) = 0.05.$$

Since A and A^c form a partition of the sample space, by the law of total probability,

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c) \\ &= 0.01 \times 0.98 + 0.99 \times 0.05 = 0.0593.\end{aligned}$$

(2) Since A and A^c form a partition of the sample space, by Bayes' theorem,

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c)} \\ &= \frac{0.01 \times 0.98}{0.01 \times 0.98 + 0.99 \times 0.05} = \frac{0.0098}{0.0593} = 0.1653.\end{aligned}$$

3. **(1)** Let A be the event that a randomly selected cadet is male, and B the event that the cadet scored full marks. Then $\mathbb{P}(A) = 0.9$, $\mathbb{P}(A^c) = 0.1$, $\mathbb{P}(B \mid A) = 0.20$, and $\mathbb{P}(B \mid A^c) = 0.21$. By the law of total probability,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c) = (0.20)(0.9) + (0.21)(0.1) = 0.201.$$

(2) By Bayes' theorem,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{(0.20)(0.9)}{0.201} \approx 0.8955.$$

4. **(1)** Let A be the event that the attack was conducted by Group A, B the event that it was conducted by Group B, and C the event that signs of unauthorized access appear. Then $\mathbb{P}(A) = 0.40$, $\mathbb{P}(B) = 0.60$, $\mathbb{P}(C \mid A) = 0.80$, and $\mathbb{P}(C \mid B) = 0.30$. Since A and B form a partition of the sample space,

$$\mathbb{P}(C) = \mathbb{P}(C \mid A)\mathbb{P}(A) + \mathbb{P}(C \mid B)\mathbb{P}(B) = (0.80)(0.40) + (0.30)(0.60) = 0.32 + 0.18 = 0.50.$$

(2) By Bayes' theorem,

$$\mathbb{P}(A \mid C) = \frac{\mathbb{P}(C \mid A)\mathbb{P}(A)}{\mathbb{P}(C)} = \frac{(0.80)(0.40)}{0.50} = \frac{0.32}{0.50} = 0.64.$$

5. **(1)** Let A be the event that an issued firearm is defective and B the event of a malfunction during shooting. Then $\mathbb{P}(A) = 0.003$, $\mathbb{P}(A^c) = 0.997$, $\mathbb{P}(B \mid A) = 0.94$, $\mathbb{P}(B \mid A^c) = 0.02$, so

$$\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c) = (0.003)(0.94) + (0.997)(0.02) = 0.0228.$$

(2) By Bayes' theorem, $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(B)} = \dfrac{(0.003)(0.94)}{0.0228} \approx 0.1237.$

6. **(1)** Let A be the event that the battery pack is defective and B the event that the radio does not work properly. Then $\mathbb{P}(A) = 0.01$, $\mathbb{P}(A^c) = 0.99$, $\mathbb{P}(B \mid A) = 0.75$, $\mathbb{P}(B \mid A^c) = 0.02$, so

$$\mathbb{P}(B) = (0.01)(0.75) + (0.99)(0.02) = 0.0075 + 0.0198 = 0.0273.$$

(2) By Bayes' theorem, $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(B)} = \dfrac{(0.01)(0.75)}{0.0273} \approx 0.2747.$

Chapter 4

§4.1 Defining random variables

1. **(1)** $\sum_x f(x) = c(1+4+9) = 14c = 1$, so $c = \dfrac{1}{14}$.

   (2) $E(X) = \dfrac{1(1)+2(4)+3(9)}{14} = \dfrac{36}{14} = \dfrac{18}{7} \approx 2.5714$ and $E(X^2) = \dfrac{1(1)+4(4)+9(9)}{14} = \dfrac{98}{14} = 7$, so $\mathrm{Var}(X) = 7 - \left(\dfrac{18}{7}\right)^2 = \dfrac{19}{49} \approx 0.3878$.

   (3) $\mathbb{P}(X > 1) = f(2) + f(3) = \dfrac{4+9}{14} = \dfrac{13}{14} \approx 0.9286$.

2. **(1)** $f(x) = 2e^{-2x} \geq 0$ for $x > 0$, and $\displaystyle\int_0^\infty 2e^{-2x}\,dx = \Big[-e^{-2x}\Big]_0^\infty = 1$, so f is a valid PDF.

   (2) $\mathbb{P}(X > 1) = \displaystyle\int_1^\infty 2e^{-2x}\,dx = e^{-2} \approx 0.1353$ and $\mathbb{P}(0.5 < X < 2) = e^{-1} - e^{-4} \approx 0.3496$.

   (3) For $x > 0$, $F(x) = \displaystyle\int_0^x 2e^{-2t}\,dt = 1 - e^{-2x}$; and $F(x) = 0$ for $x \leq 0$.

3. **(1)** All probabilities are nonnegative and $0.1 + 0.3 + 0.4 + 0.2 = 1$, so it is a valid PMF.

   (2) $\mathbb{P}(X \geq 2) = f(2) + f(3) = 0.4 + 0.2 = 0.6$.

   (3) $F(x) = 0$ for $x < 0$; 0.1 for $0 \leq x < 1$; 0.4 for $1 \leq x < 2$; 0.8 for $2 \leq x < 3$; and 1 for $x \geq 3$. Then $\mathbb{P}(1 < X \leq 3) = F(3) - F(1) = 1 - 0.4 = 0.6$.

§ 4.2 Expectation and variance

1. **(1)** $E(2X+5) = 2E(X) + 5 = 2(4) + 5 = 13$. **(2)** $\mathrm{Var}(2X+5) = 2^2\mathrm{Var}(X) = 4(9) = 36$. **(3)** $\mathrm{SD}(3 - X) = |-1|\mathrm{SD}(X) = \sqrt{9} = 3$.

2. $E(X) = (-1)(0.5) + 0(0.3) + 2(0.2) = -0.1$ and $E(X^2) = 1(0.5) + 0(0.3) + 4(0.2) = 1.3$, so $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = 1.3 - (-0.1)^2 = 1.29$.

3. **(1)** The face value has mean $\dfrac{1+2+\cdots+6}{6} = 3.5$, so the expected net gain is $3.5 - 2 = 1.5$ dollars per play. **(2)** The game is favorable: the expected net gain is positive (+\$1.5), so on average the player wins.

§ 4.3 Joint distributions

1. **(1)** Using $f_X(x) = \sum_{y=0}^{1} f(x,y)$, we get $f_X(0) = \frac{1}{9}$, $f_X(1) = \frac{3}{9}$, $f_X(2) = \frac{5}{9}$:

$$f_X(x) = \begin{cases} 1/9, & x = 0 \\ 3/9, & x = 1 \\ 5/9, & x = 2 \\ 0, & \text{otherwise.} \end{cases}$$

   (2) Using $\mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2$ with $E(Y) = \frac{2}{3}$,

$$E(Y^2) = \sum_{y=0}^{1} y^2 f_Y(y) = 0^2 \cdot \tfrac{1}{3} + 1^2 \cdot \tfrac{2}{3} = \tfrac{2}{3}.$$

Therefore,

$$\mathrm{Var}(Y) = E(Y^2) - \left[E(Y)\right]^2 = \frac{2}{3} - \left(\frac{2}{3}\right)^2 = \frac{2}{3} - \frac{4}{9} = \frac{6}{9} - \frac{4}{9} = \frac{2}{9}.$$

(3) First compute $E(X)$:

$$E(X) = \sum_{x=0}^{2} x\, f_X(x) = 0 \times \tfrac{1}{9} + 1 \times \tfrac{1}{3} + 2 \times \tfrac{5}{9} = 0 + \tfrac{3}{9} + \tfrac{10}{9} = \tfrac{13}{9}.$$

Using the linearity of expectation and $E(Y) = \frac{2}{3}$,

$$\begin{aligned} E(2X - 3Y + 4) &= 2\,E(X) - 3\,E(Y) + 4 \\ &= 2 \times \frac{13}{9} - 3 \times \frac{2}{3} + 4 = \frac{26}{9} - \frac{18}{9} + \frac{36}{9} = \frac{44}{9}. \end{aligned}$$

(4) Check whether $f(x, y) = f_X(x)\, f_Y(y)$ holds for all (x, y). For $(x, y) = (0, 0)$,

$$f(0,0) = 0, \qquad f_X(0) \cdot f_Y(0) = \frac{1}{9} \times \frac{1}{3} = \frac{1}{27}.$$

Since $0 \neq \dfrac{1}{27}$, we have $f(0,0) \neq f_X(0)\, f_Y(0)$, so X and Y are **not independent**.

2. **(1)**

$$f_X(x) = \int_0^1 e^{-x}\, dy = e^{-x}, \quad x > 0.$$

(2) Since $f(x, y) = f_X(x) f_Y(y)$, the random variables X and Y are independent.

(3)

$$E(Y) = \int_0^1 y \cdot 1\, dy = \int_0^1 y\, dy = \tfrac{1}{2}.$$

(4) Method 1: Since X and Y are independent, and $E(X)$, $E(Y)$, and $E(XY)$ exist, $\mathrm{Cov}(X, Y) = 0$.
Method 2:

$$\mathrm{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) = \tfrac{1}{2} - 1 \cdot \tfrac{1}{2} = 0.$$

3. **(1)** Summing over the other variable,

$$f_X(x) = \sum_{y=1}^{2} \frac{x^2 y}{42} = \frac{1}{14} x^2, \quad x = 1, 2, 3; \qquad f_Y(y) = \sum_{x=1}^{3} \frac{x^2 y}{42} = \frac{1}{3} y, \quad y = 1, 2.$$

(2) Since $f(x, y) = f_X(x)\, f_Y(y)$ for all (x, y), the random variables X and Y are **independent**.

(3) $E(Y) = \displaystyle\sum_{y=1}^{2} y\, f_Y(y) = \sum_{y=1}^{2} \frac{y^2}{3} = \frac{1+4}{3} = \frac{5}{3}.$

(4) $E(Y^2) = \sum_{y=1}^{2} y^2 f_Y(y) = \sum_{y=1}^{2} \frac{y^3}{3} = \frac{1+8}{3} = 3$, so

$$\mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2 = 3 - \frac{25}{9} = \frac{2}{9}.$$

4. **(1)** Integrating the constant density over the region where $x < y$ (which requires $35 \le x < y \le 40$),

$$\mathbb{P}(X < Y) = \int_{35}^{40}\int_{35}^{y} \frac{1}{200}\,dx\,dy = \int_{35}^{40} \frac{y-35}{200}\,dy = \left.\frac{(y-35)^2}{400}\right|_{35}^{40} = \frac{25}{400} = 0.0625.$$

(2) Let W be the number of times (out of 30) that Alex arrives earlier. Then $W \sim B(30, 0.0625)$, so

$$E(W) = 30 \times 0.0625 = 1.875.$$

5. **(1)** $f_Y(y) = \int_0^1 4xy\,dx = 2y$ for $0 < y < 1$ (and 0 otherwise).

(2) $E(X) = \int_0^1 x\,f_X(x)\,dx = \int_0^1 2x^2\,dx = \frac{2}{3}$.

(3) Since $f(x,y) = 4xy = (2x)(2y) = f_X(x)f_Y(y)$ for all x, y, the random variables X and Y are independent.

Chapter 5

§5.1 Uniform distribution

1. **(1)** Since X is uniformly distributed on the interval $(0, 15)$, its probability density function is

$$f(x) = \begin{cases} \frac{1}{15} & 0 < x < 15, \\ 0 & \text{otherwise.} \end{cases}$$

(2)

$$\mathbb{P}(X < 3) = \int_{-\infty}^{3} f(x)dx = \int_0^3 \frac{1}{15}dx = \frac{3}{15} = 0.2.$$

(3)

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{15} x \cdot \frac{1}{15}dx = \frac{15}{2} = 7.5.$$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-7.5)^2 f(x)dx = \int_0^{15} (x-7.5)^2 \cdot \frac{1}{15}dx = \frac{1}{15}\left[\frac{1}{3}(x-7.5)^3\right]_0^{15} = \frac{2\times 7.5^3}{15\times 3} = 18.75.$$

In general, if X is uniformly distributed in (a, b), then $E(X) = \frac{a+b}{2}$ and $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

2. **(1)** Since X follows a uniform distribution on $(0, 10)$, $f_X(x) = \begin{cases} 1/10, & 0 < x < 10, \\ 0, & \text{otherwise.} \end{cases}$ By definition of the mean, $E(X) = \int_0^{10} x \cdot \frac{1}{10}\,dx = 5$.

(2) $\mathbb{P}(Y = 10) = \mathbb{P}(X < 4) = 4/10$ and $\mathbb{P}(Y = 5) = \mathbb{P}(4 \le X < 10) = 6/10$.

(3) By definition of the mean,

$$E(Y) = \sum_{\forall y} y\, f(y) = 10 \times \frac{4}{10} + 5 \times \frac{6}{10} = 7.$$

3. **(1)** For $0 < x < 30$,

$$F_X(x) = \mathbb{P}(X \le x) = \int_0^x \frac{1}{30}\, dt = \frac{x}{30}.$$

(2) For a single device, $\mathbb{P}(X > 10) = 1 - F_X(10) = 1 - \frac{10}{30} = \frac{2}{3}$. Let Y be the number of the five devices lasting more than 10 hours, so $Y \sim B(5, 2/3)$. Then

$$\mathbb{P}(Y \ge 2) = 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) = 1 - \left(\tfrac{1}{3}\right)^5 - 5 \cdot \tfrac{2}{3}\left(\tfrac{1}{3}\right)^4 = 1 - \frac{1}{243} - \frac{10}{243} = \frac{232}{243} \approx 0.9547.$$

§ 5.2 Normal distribution

1. Let X denote the battery life (in hours) of a portable radio produced by the company.

The population follows a normal distribution with mean $\mu = 45$ and standard deviation $\sigma = 3$, so $X \sim N(45, 3^2)$.

For a random sample of size $n = 16$,

$$E(\overline{X}) = \mu = 45, \quad \mathrm{Var}(\overline{X}) = \frac{3^2}{16} = \left(\frac{3}{4}\right)^2,$$

so $\overline{X} \sim N\left(45, 0.75^2\right)$. Standardizing, $Z = \dfrac{\overline{X} - 45}{0.75} \sim N(0, 1)$. Therefore,

$$\begin{aligned}
\mathbb{P}(44 < \overline{X} < 47) &= \mathbb{P}\left(\frac{44 - 45}{0.75} < \frac{\overline{X} - 45}{0.75} < \frac{47 - 45}{0.75}\right) \\
&= \mathbb{P}(-1.3333 < Z < 2.6667) \\
&= \mathbb{P}(Z < 2.6667) - \mathbb{P}(Z < -1.3333) \\
&= \mathbb{P}(Z < 2.6667) - \left(1 - \mathbb{P}(Z < 1.3333)\right) \\
&= 0.9962 - (1 - 0.9088) = 0.9050.
\end{aligned}$$

2. We have $X \sim N(100, 3^2)$ and $Y \sim N(95, 4^2)$.

We wish to find $\mathbb{P}(X \ge Y + 17.5) = \mathbb{P}(X - Y \ge 17.5)$. Consider $X - Y$:

$$E(X - Y) = E(X) - E(Y) = 100 - 95 = 5.$$

Since X and Y are independent,

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 3^2 + 4^2 = 25 = 5^2.$$

Therefore $X - Y \sim N(5, 5^2)$. Standardizing, $Z = \dfrac{(X-Y)-5}{5} \sim N(0,1)$. Thus,

$$\mathbb{P}(X - Y \geq 17.5) = \mathbb{P}\left(\frac{(X-Y)-5}{5} \geq \frac{17.5-5}{5}\right) = \mathbb{P}\left(Z \geq \frac{12.5}{5}\right)$$
$$= \mathbb{P}(Z \geq 2.5) = 1 - \mathbb{P}(Z < 2.5) = 1 - 0.9938 = 0.0062.$$

3. **(1)** Let the random variable X be the temperature (in °F) on a randomly selected day in June. Then
$$X \sim N(77, 5^2)$$
The probability of observing an 83°F temperature or higher is:
$$\mathbb{P}(X \geq 83) = \mathbb{P}\left(\frac{X-77}{5} \geq \frac{83-77}{5}\right) = \mathbb{P}(Z \geq 1.2) \approx 1 - 0.8849 = 0.1151$$
(2) We want the 10th lower percentile of $X \sim N(77, 5^2)$. That is, find x such that $\mathbb{P}(X \leq x) = 0.10$.
$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X-77}{5} \leq \frac{x-77}{5}\right) = \mathbb{P}\left(Z \leq \frac{x-77}{5}\right) = 0.10.$$
Since this the left-tail area of standard normal distribution, $\dfrac{x-77}{5} = -z_{0.10} \approx -1.28$
Transform back to the original scale:
$$x = 77 + (-1.28)(5) = 70.6.$$

4. **(1)** Since $\mathbb{P}(X \geq 5) = 0.5$ and the normal distribution is symmetric about its mean, $E(X) = \mu = 5$. Standardizing $\mathbb{P}(X \geq 8.92) = 0.025$,
$$\mathbb{P}\left(Z \geq \frac{8.92-5}{\sigma}\right) = 0.025 = \mathbb{P}(Z \geq z_{0.025}), \quad \text{so} \quad \frac{3.92}{\sigma} = z_{0.025} = 1.96,$$
giving $\sigma = \dfrac{3.92}{1.96} = 2$ and $\text{Var}(X) = \sigma^2 = 4$.
(2) We seek x with $\mathbb{P}(X < x) = 0.025$. Standardizing, $\dfrac{x-5}{2} = -z_{0.025} = -1.96$, so $x = 5 + (-1.96)(2) = 1.08$. (Equivalently, by symmetry $\frac{1}{2}(x + 8.92) = 5$ gives $x = 1.08$.)

5. **(1)** Let $X \sim N(60, 10^2)$ be the midterm score and $Y \sim N(70, 5^2)$ the final score. The final grade is $Z = 0.4X + 0.6Y$, with
$$E(Z) = 0.4(60) + 0.6(70) = 66, \qquad \text{Var}(Z) = 0.4^2(100) + 0.6^2(25) = 16 + 9 = 25,$$
so $Z \sim N(66, 5^2)$.
(2) This cadet's grade is $0.4(70) + 0.6(80) = 76$. Standardizing,
$$\mathbb{P}(Z \geq 76) = \mathbb{P}\left(\frac{Z-66}{5} \geq \frac{76-66}{5}\right) = \mathbb{P}(Z \geq 2) = 0.0228,$$
so the cadet is in the top 2.28% (about the 97.72th percentile).

(3) The cutoff c satisfies $\mathbb{P}(Z > c) = 0.20$, i.e. $\dfrac{c-66}{5} = z_{0.20} \approx 0.8416$, so $c = 66 + 0.8416(5) \approx 70.2081$. A student needs a final grade of at least about 70.21 to receive an A.

6. **(1)** Since $W = 0.3Q + 0.3M + 0.4F$ is a linear combination of independent normals,

$$E(W) = 0.3(80) + 0.3(60) + 0.4(70) = 70,$$

$$\mathrm{Var}(W) = 0.3^2(5^2) + 0.3^2(20^2) + 0.4^2(5^2) = 2.25 + 36 + 4 = 42.25,$$

so $W \sim N(70,\, 6.5^2)$ (since $\sqrt{42.25} = 6.5$).

(2) This student's score is $W = 0.3(90) + 0.3(80) + 0.4(85) = 85$. Standardizing, $Z = \dfrac{85-70}{6.5} \approx 2.3077$, so

$$\mathbb{P}(W > 85) = \mathbb{P}(Z > 2.3077) \approx 0.0105,$$

i.e. about 1.05% of students scored higher.

7. **(1)** Let $X \sim N(30, 3^2)$ be the completion time. Then

$$\mathbb{P}(27 \le X \le 36) = \mathbb{P}(-1 \le Z \le 2) = 1 - \mathbb{P}(Z \le -2) - \mathbb{P}(Z \le -1) \approx 1 - 0.0228 - 0.1587 = 0.8185.$$

(2) The top-20% cutoff c satisfies $\mathbb{P}(X \ge c) = 0.20$, i.e. $\dfrac{c-30}{3} = z_{0.20} \approx 0.8416$, so $c = 30 + 0.8416(3) \approx 32.5248$ minutes.

§ 5.3 Binomial distribution

1. **(1)** Since X is the number of license holders in $n = 10$ trials with success probability $p = 0.6$, we have $X \sim B(10,\, 0.6)$. Therefore,

$$\mathbb{P}(X = 8) + \mathbb{P}(X = 9) = \binom{10}{8}(0.6)^8(0.4)^2 + \binom{10}{9}(0.6)^9(0.4)^1 \approx 0.1612.$$

(2) For $X \sim B(n, p)$, we have $E(X) = np$ and $\mathrm{Var}(X) = np(1-p)$. Since $X \sim B(10,\, 0.6)$,

$$E(X) = 10 \times 0.6 = 6, \qquad \mathrm{Var}(X) = 10 \times 0.6 \times 0.4 = 2.4.$$

2. **(1)** Since $X \sim B\left(3, \frac{1}{3}\right)$,

$$f_X(x) = \begin{cases} \binom{3}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{3-x}, & x = 0, 1, 2, 3, \\ 0, & \text{otherwise.} \end{cases}$$

(2) (1) $\mathbb{P}(X = 0, Y = 3) = \left(\dfrac{2}{3}\right)^3 = \dfrac{8}{27}$ (2) $\mathbb{P}(X = 3, Y = 1) = 0$

(3) Note that $f(1,1) = 0$, but $f_X(1)\, f_Y(1) = \dfrac{12}{27} \times \dfrac{6}{27} = \dfrac{8}{81} \neq 0$.
Since $f(1,1) \neq f_X(1)\, f_Y(1)$, X and Y are **not independent**.
(4) [Method 1] Since $X \sim B(3, \frac{1}{3})$ and $Y \sim B(3, \frac{2}{3})$,

$$E(X) = 1, \quad E(Y) = 2, \quad \mathrm{Var}(X) = \frac{2}{3}, \quad \mathrm{Var}(Y) = \frac{2}{3}.$$

$$E(XY) = \sum_{x=0}^{3}\sum_{y=0}^{3} xy\, f(x,y) = 1\times 2\times\frac{12}{27} + 2\times 1\times\frac{6}{27} = \frac{4}{3},$$

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{4}{3} - 2 = -\frac{2}{3}.$$

$$\therefore\ \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y) = \frac{2}{3} + \frac{2}{3} - 2\left(-\frac{2}{3}\right) = \frac{8}{3}.$$

[Method 2] Since $Y = 3 - X$,

$$\mathrm{Var}(X-Y) = \mathrm{Var}(2X-3) = 4\mathrm{Var}(X) = 4np(1-p) = 4\times 3\times\frac{1}{3}\times\frac{2}{3} = \frac{8}{3}.$$

3. **(1)** The number who had chickenpox is $X \sim B(100, 0.90)$, so

$$\mathbb{P}(X = 97) = \binom{100}{97}(0.90)^{97}(0.10)^3 \approx 0.0059.$$

(2) Let $Y \sim B(10, 0.90)$. Then

$$\mathbb{P}(Y \geq 1) = 1 - \mathbb{P}(Y = 0) = 1 - (0.10)^{10} \approx 1.0000.$$

4. **(1)** Since $X \sim B(144, 0.90)$,

$$\mathbb{P}(X = 2)\times 10^{142} = \binom{144}{2}(0.90)^2(0.10)^{142}\times 10^{142} = \binom{144}{2}(0.81) \approx 8339.76.$$

(2) $E(X) = np = 144(0.90) = 129.6$ and $\mathrm{Var}(X) = np(1-p) = 144(0.90)(0.10) = 12.96\,(= 3.6^2)$.
(3) By the normal approximation, $X \dot{\sim} N(129.6,\ 3.6^2)$, so

$$\mathbb{P}(X \geq 135) = \mathbb{P}\left(Z \geq \frac{135 - 129.6}{3.6}\right) = \mathbb{P}(Z \geq 1.5) \approx 0.0668.$$

5. **(1)** $X \sim B(10, 0.1)$, so $E(X) = 10(0.1) = 1$ and $\mathrm{Var}(X) = 10(0.1)(0.9) = 0.9$.
(2) $E(Y) = \frac{1}{10}E(X) = 0.1$ and $\mathrm{Var}(Y) = \frac{1}{10^2}\mathrm{Var}(X) = 0.009$.
(3)

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - (0.9)^{10} - 10(0.1)(0.9)^9 = 1 - 0.3487 - 0.3874 = 0.2639.$$

Chapter 6

§6.1 Point estimates and sampling variability

1. **(1)** For the sample proportion $\hat{p}$,

$$E(\hat{p}) = 0.2, \quad \mathrm{Var}(\hat{p}) = \frac{0.2\times 0.8}{60} = 0.0027.$$

Since $n = 60$ is large and $np = 12 > 10$, $n(1-p) = 48 > 10$, by the Central Limit Theorem,

$$\hat{p} \dot{\sim} N(0.2,\ 0.0027).$$

(2) Since $\hat{p} \dot{\sim} N(0.2,\ 0.0027)$,

$$\mathbb{P}\left(\hat{p} \geq \frac{1}{6}\right) = \mathbb{P}\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{1/6 - 1/5}{\sqrt{0.0027}}\right)$$
$$\approx \mathbb{P}(Z \geq -0.6416) = \mathbb{P}(Z < 0.6416) = 0.7394.$$

2. **(a)**
$$E(\hat{p}) = p = 0.75, \quad SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.75(1-0.75)}{300}} = 0.025$$

(b) We have $\hat{p} \dot{\sim} N(0.75, 0.025^2)$.

$$\mathbb{P}(\hat{p} < 0.74) = \mathbb{P}\left(\frac{\hat{p} - 0.75}{0.025} < \frac{0.74 - 0.75}{0.025}\right) = \mathbb{P}\left(Z < -0.4\right) = 0.3446.$$

3. **(1)** Since $X \sim B(1000,\ 0.02)$,

$$\mathbb{P}(X = 0) \times 10^9 = \binom{1000}{0}(0.02)^0(0.98)^{1000} \times 10^9 \approx 1.6830.$$

(2) $E(\hat{p}) = p = 0.02$ and $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.02(0.98)}{1000}} \approx 0.0044$.

(3) The sample is random (independence), and $np = 20 \geq 10$, $n(1-p) = 980 \geq 10$, so the success–failure condition holds and $\hat{p} \dot{\sim} N(0.02,\ 0.0044^2)$. Therefore

$$\mathbb{P}(\hat{p} < 0.019) = \mathbb{P}\left(Z < \frac{0.019 - 0.02}{0.0044}\right) \approx \mathbb{P}(Z < -0.2273) = 0.4101.$$

§ 6.2 Confidence intervals for a proportion

1. **(1)** The sample is random, so the independence condition holds; and $n\hat{p} = 214 \geq 10$ and $n(1-\hat{p}) = 786 \geq 10$, so the success–failure condition holds.

(2) With $\hat{p} = 214/1000 = 0.214$, the standard error is $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.214 \times 0.786/1000} \approx 0.0130$. The 95% confidence interval is

$$0.214 \pm 1.96 \times 0.0130 = 0.214 \pm 0.0254 = (0.189,\ 0.239).$$

2. **(1)** With $\hat{p} = \frac{137}{850} \approx 0.1612$, the 95% confidence interval is

$$\hat{p} \pm z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1612 \pm 1.96\sqrt{\frac{0.1612(0.8388)}{850}} \approx 0.1612 \pm 0.0247 = (0.1365,\ 0.1859).$$

(2) (a) *smaller* (b) *smaller* (c) *larger*.

3. With $\hat{p} = \frac{440}{800} = 0.55$, the standard error is $\sqrt{\frac{0.55(1-0.55)}{800}} \approx 0.0176$, so the 95% confidence interval is

$$0.55 \pm 1.96(0.0176) \approx (0.5155,\ 0.5845).$$

Chapter 7

§7.1 Sampling distribution

1. Let X represent the lifetime of a flashlight produced at Factory K. Since the sample size is $n = 100$, which is sufficiently large, the **Central Limit Theorem** applies:

$$\overline{X} \sim N(1500, (100/\sqrt{100})^2) = N(1500, 10^2)$$

Thus, the standardized variable follows: $Z = \dfrac{\overline{X} - 1500}{10} \dot{\sim} N(0,1)$

Now, we compute:

$$\mathbb{P}(\overline{X} \geq 1510) = \mathbb{P}\left(\frac{\overline{X} - 1500}{10} \geq \frac{1510 - 1500}{10}\right)$$

$$= \mathbb{P}(Z \geq 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

2. **(1)** The possible values of $\overline{X}_2$ and their probabilities are:

$$f(x) = \begin{cases} 0.16, & x = 0 \\ 0.48, & x = 1.5 \\ 0.36, & x = 3 \\ 0, & \text{otherwise} \end{cases}$$

(2) The expectation of $\overline{X}_2$ is:

$$E(\overline{X}_2) = E(X_1) = 3 \times 0.6 + 0 \times 0.4 = 1.8.$$

The variance of X_1 is given by:

$$Var(X_1) = E(X_1^2) - [E(X_1)]^2 = 2.16.$$

Since $\overline{X}_2$ is the mean of two independent trials,

$$Var(\overline{X}_2) = \frac{Var(X_1)}{2} = \frac{2.16}{2} = 1.08.$$

(3) Since the sample size is sufficiently large ($n = 54$), we apply the Central Limit Theorem:

$$\overline{X}_{54} \dot{\sim} N\left(1.8, \frac{2.16}{54}\right).$$

Standardizing:

$$Z = \frac{\overline{X}_{54} - 1.8}{\sqrt{2.16/54}} \dot{\sim} N(0,1).$$

Thus,

$$\mathbb{P}(\overline{X}_{54} \geq 2) = \mathbb{P}\left(\frac{\overline{X}_{54} - 1.8}{\sqrt{2.16/54}} \geq \frac{2 - 1.8}{\sqrt{2.16/54}}\right) = \mathbb{P}(Z \geq 1).$$

From standard normal tables, $\mathbb{P}(Z \geq 1) = 0.1587$.

3. **(1)** Since $X_1, X_2 \sim N(200, 200)$ independently,

$$E(\overline{X}) = \frac{E(X_1) + E(X_2)}{2} = 200, \qquad \mathrm{Var}(\overline{X}) = \frac{1}{4}\Big(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\Big) = \frac{1}{4}(200 + 200) = 100.$$

(2) Thus $\overline{X} \sim N(200, 100)$, i.e. $\mathrm{SD}(\overline{X}) = 10$, so

$$\mathbb{P}(\overline{X} < 180) = \mathbb{P}\Big(Z < \frac{180 - 200}{10}\Big) = \mathbb{P}(Z < -2) = 0.0228.$$

4. **(1)** Collecting the joint probabilities by the value of $\overline{x} = \frac{x_1+x_2}{2}$,

$$\mathbb{P}(\overline{X} = 1) = 0.36,\ \mathbb{P}(\overline{X} = 1.5) = 0.24,\ \mathbb{P}(\overline{X} = 2) = 0.28,\ \mathbb{P}(\overline{X} = 2.5) = 0.08,\ \mathbb{P}(\overline{X} = 3) = 0.04.$$

(2) With $n = 2$ and $\overline{X} = \frac{X_1+X_2}{2}$,

$$S^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = \left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2 = \frac{1}{2}(X_1 - X_2)^2.$$

(3) Note $\mathbb{P}(\overline{X} = 1, S^2 = 0) = \mathbb{P}(X_1 = 1, X_2 = 1) = 0.36$, whereas $\mathbb{P}(\overline{X} = 1) = 0.36$ and $\mathbb{P}(S^2 = 0) = 0.44$, so

$$\mathbb{P}(\overline{X} = 1,\ S^2 = 0) = 0.36 \neq 0.36 \times 0.44 = \mathbb{P}(\overline{X} = 1)\,\mathbb{P}(S^2 = 0).$$

Therefore $\overline{X}$ and S^2 are **not independent**.

§ 7.2 Point estimation

1. **(1)**

$$E(\hat{\mu}_1) = \frac{1}{3}(E(X_1) + E(X_2) + E(X_3)) = \frac{1}{3}(\mu + \mu + \mu) = \mu,$$

$$E(\hat{\mu}_2) = \frac{1}{4}(E(X_1) + 2E(X_2) + E(X_3)) = \frac{1}{4}(\mu + 2\mu + \mu) = \mu,$$

$$E(\hat{\mu}_3) = E(X_3) = \mu.$$

Therefore $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ are all unbiased estimators of μ.

(2) Since all three estimators are unbiased, we compare their variances:

$$\mathrm{Var}(\hat{\mu}_1) = \frac{\sigma^2}{3},$$

$$\mathrm{Var}(\hat{\mu}_2) = \frac{1}{16}\Big(\mathrm{Var}(X_1) + 4\,\mathrm{Var}(X_2) + \mathrm{Var}(X_3)\Big) = \frac{1}{16}(\sigma^2 + 4\sigma^2 + \sigma^2) = \frac{6\sigma^2}{16} = \frac{3\sigma^2}{8},$$

$$\mathrm{Var}(\hat{\mu}_3) = \sigma^2.$$

Since $\text{Var}(\hat{\mu}_1) < \text{Var}(\hat{\mu}_2) < \text{Var}(\hat{\mu}_3)$, the most efficient estimator of μ is $\hat{\mu}_1$.

2.
$$E(\hat{\mu}) = E\left(\frac{2X_1 + X_2 + 2X_3 + X_4 + X_5}{7}\right)$$
$$= \frac{1}{7}E(2X_1 + X_2 + 2X_3 + X_4 + X_5)$$
$$= \frac{1}{7}(2\mu + \mu + 2\mu + \mu + \mu) = \mu$$

Since $E(\hat{\mu}) = \mu$, the estimator $\hat{\mu}$ is an unbiased estimator of μ.

3. **(1)** $E(\hat{p}_1) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = p$ and $E(\hat{p}_2) = \frac{1}{2}\big(E(X_1) + E(X_n)\big) = p$, so both are **unbiased**.

(2) Their variances are $\text{Var}(\hat{p}_1) = \frac{p(1-p)}{n}$ and $\text{Var}(\hat{p}_2) = \frac{p(1-p)}{2}$. Since $\text{Var}(\hat{p}_1) < \text{Var}(\hat{p}_2)$ for $n > 2$, the estimator $\hat{p}_1$ is **more efficient**.

(3) Since $\lim_{n\to\infty} E(\hat{p}_1) = p$ and $\lim_{n\to\infty} \text{Var}(\hat{p}_1) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0$, the estimator $\hat{p}_1$ is **consistent**.

(4) Writing $\hat{p} = \hat{p}_1$, by the Central Limit Theorem $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1)$ for large n, giving the 95% confidence interval
$$\left[\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

4. **(1)** Using $E(X_i^2) = \frac{\theta}{2}$,
$$E(\hat{\theta}) = \frac{k}{n}\sum_{i=1}^{n} E(X_i^2) = \frac{k}{n} \cdot n \cdot \frac{\theta}{2} = \frac{k\theta}{2}.$$

Setting $E(\hat{\theta}) = \theta$ gives $\frac{k\theta}{2} = \theta$, so $k = 2$.

(2) With $k = 2$, $\hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i^2$ is unbiased, so $\lim_{n\to\infty} E(\hat{\theta}) = \theta$. Moreover, by independence,
$$\text{Var}(\hat{\theta}) = \frac{4}{n^2}\sum_{i=1}^{n} \text{Var}(X_i^2) = \frac{4}{n}\text{Var}(X_1^2) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n} \longrightarrow 0.$$

Since $\lim_{n\to\infty} E(\hat{\theta}) = \theta$ and $\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0$, $\hat{\theta}$ is **consistent**.

5. **(1)** Since $E(X_1) = \int_0^{\theta} x\,\frac{3x^2}{\theta^3}\,dx = \frac{3}{4}\theta$,
$$E(\hat{\theta}) = \frac{4}{3}E(\overline{X}_n) = \frac{4}{3}E(X_1) = \frac{4}{3} \cdot \frac{3}{4}\theta = \theta,$$

so $\hat{\theta}$ is unbiased.

(2) Also $E(X_1^2) = \displaystyle\int_0^\theta x^2 \frac{3x^2}{\theta^3}\,dx = \frac{3}{5}\theta^2$, so $\mathrm{Var}(X_1) = \frac{3}{5}\theta^2 - \left(\frac{3}{4}\theta\right)^2 = \frac{3}{80}\theta^2 < \infty$. Hence

$$\mathrm{Var}(\hat{\theta}) = \frac{16}{9}\mathrm{Var}(\overline{X}_n) = \frac{16}{9n}\mathrm{Var}(X_1) = \frac{16}{9n}\cdot\frac{3}{80}\theta^2 = \frac{\theta^2}{15n} \longrightarrow 0,$$

and since $E(\hat{\theta}) = \theta$, the estimator $\hat{\theta}$ is consistent.

(3) For large n, $\dfrac{\hat{\theta}-\theta}{\theta/\sqrt{15n}} \dot{\sim} N(0,1)$. With $n = 60$ this is $\dfrac{\hat{\theta}-\theta}{\theta/30}$, so

$$1-\alpha \approx \mathbb{P}\left(\frac{\hat{\theta}}{1+z_{\alpha/2}/30} \le \theta \le \frac{\hat{\theta}}{1-z_{\alpha/2}/30}\right).$$

With $\hat{\theta} = \frac{4}{3}\overline{x} = 4$ and $z_{0.025} = 1.96$, a 95% confidence interval for θ is approximately $[\,3.7547,\ 4.2796\,]$.

Chapter 8

§8.1 Confidence interval for a population mean

1. To construct a 90% confidence interval for mean μ, we use the t-distribution. The degrees of freedom are $df = n-1 = 14$. Therefore, the 90% confidence interval is:

$$\begin{aligned}\overline{x} \pm t_{0.05}(14)\cdot\frac{s}{\sqrt{n}} &= 299.16 \pm 1.7613\cdot\frac{4.44}{\sqrt{15}}\\ &\approx 299.16 \pm 2.0192 = (297.1408, 301.1792)\end{aligned}$$

2. **(1)** When σ is known, a 95% confidence interval is $\overline{x} \pm z_{0.025}\dfrac{\sigma}{\sqrt{n}}$. With $\overline{x} = 97$, $\sigma = 2$, $n = 25$, $z_{0.025} = 1.96$,

$$97 \pm 1.96\cdot\frac{2}{\sqrt{25}} \approx (96.2160,\ 97.7840).$$

(2) When σ is unknown, a 90% confidence interval is $\overline{x} \pm t_{0.05}(n-1)\dfrac{s}{\sqrt{n}}$. With $s = 2.0616$, $n = 25$, $t_{0.05}(24) = 1.7109$,

$$97 \pm 1.7109\cdot\frac{2.0616}{\sqrt{25}} \approx (96.2945,\ 97.7055).$$

3. Since σ is unknown, we use the t-distribution with $df = n-1 = 9$. With $t_{0.025}(9) = 2.2622$,

$$\overline{x} \pm t_{0.025}(9)\cdot\frac{s}{\sqrt{n}} = 58 \pm 2.2622\cdot\frac{8}{\sqrt{10}} \approx 58 \pm 5.7230 = (52.2770,\ 63.7230).$$

4. **(1)** smaller (the variance of $\overline{X}$ is σ^2/n). **(2)** symmetric (about 0). **(3)** thicker (heavier tails than Z, from estimating σ). **(4)** thinner (the t curve approaches Z as $df \to \infty$).

§8.2 Hypothesis test for a population mean

1. **(1)**
$$H_0 : \mu = 655, \qquad H_A : \mu < 655.$$
(2) Since the population variance σ^2 is unknown, we use the T statistic. When H_0 is true,
$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 655}{S/\sqrt{9}} = \frac{\overline{X} - 655}{S/3} \sim t(8).$$
(3)
$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{5832}{9} = 648, \qquad s = \sqrt{\frac{1}{9-1}\sum_{i=1}^{9}(x_i - \bar{x})^2} = \sqrt{\frac{3528}{8}} = 21.$$
The observed test statistic is
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{648 - 655}{21/\sqrt{9}} = \frac{-7}{7} = -1.$$
(4) Since $H_A : \mu < 655$ is a one-sided (left-tailed) test,
$$p\text{-value} = \mathbb{P}(T < t_0^*) = \mathbb{P}(T < -1) = 0.1733.$$
Since p-value $= 0.1733 > \alpha = 0.05$, we fail to reject H_0. There is insufficient evidence at $\alpha = 0.05$ to support the claim that the average breaking strength of the steel components is less than 655 MPa.

2. **(1)** Since $\sigma^2 = 4$ (i.e., $\sigma = 2$) is known and $n = 25$, $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X}-24}{2/\sqrt{25}} = \frac{\overline{X}-24}{0.4} \sim N(0, 1)$.
(2) Since this is a two-sided test, the rejection region is
$$|z| > z_{0.025} = 1.96.$$
(3) The observed test statistic is
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{25 - 24}{0.4} = 2.5.$$
Since $|z| = 2.5 > 1.96$, we reject H_0. There is sufficient evidence at $\alpha = 0.05$ to conclude that the average head size of recruits differs from 24 inches.

3. **(1)** $H_0 : \mu = 12.7$ versus $H_A : \mu \neq 12.7$.
(2) Since σ^2 is unknown, under H_0,
$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 12.7}{S/\sqrt{10}} \sim t(9).$$
(3) $t = \dfrac{12.744 - 12.7}{0.0783/\sqrt{10}} = 1.7770.$
(4) Since the test is two-sided,
$$p\text{-value} = 2\mathbb{P}(T > 1.7770) = 2(0.0546) = 0.1092.$$

Since $0.1092 > \alpha = 0.05$, we fail to reject H_0. There is insufficient evidence at $\alpha = 0.05$ to conclude that the average diameter differs from 12.7 mm.

4. **(1)** $H_0 : \mu = 20$ versus $H_A : \mu < 20$.

   (2) Since σ is known, under H_0, $Z = \dfrac{\overline{X} - 20}{4/\sqrt{16}} \sim N(0,1)$.

   (3) This is a left-tailed test, so the rejection region is $z < -z_{0.05} = -1.6449$.

   (4) $z = \dfrac{18.2 - 20}{4/\sqrt{16}} = -1.8$.

   (5) Since $z = -1.8 < -1.6449$, we reject H_0. There is sufficient evidence at $\alpha = 0.05$ that the new supplement gives a smaller average recovery time (less than 20 days).

5. **(1)** $H_0 : \mu = 16$ versus $H_A : \mu < 16$.

   (2) Since σ is unknown, under H_0

   $$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 16}{S/\sqrt{9}} \sim t(8).$$

   (3) $t = \dfrac{15.84 - 16}{0.24/\sqrt{9}} = \dfrac{-0.16}{0.08} = -2$.

   (4) p-value $= \mathbb{P}(T < -2) \approx 0.0403$ for $T \sim t(8)$. Since $0.0403 < 0.05$, we reject H_0. There is statistically significant evidence that the average storage capacity of Company A's USB drives is less than 16 GB.

§ 8.3 Power calculations for a population mean

1. With the rejection region $\overline{x} < c = 28.355$ and $SE = 1$ unchanged, suppose the true mean is $\mu = 28$, so $\overline{X} \sim N(28, 1^2)$. Then

   $$\text{Power} = \mathbb{P}(\overline{X} < 28.355 \mid \mu = 28) = \mathbb{P}\left(Z < \frac{28.355 - 28}{1}\right) = \mathbb{P}(Z < 0.355) \approx 0.639.$$

   This is lower than the power of about 0.91 at $\mu = 27$ because $\mu = 28$ is closer to the null value $\mu_0 = 30$. The effect to be detected is smaller ($\Delta = 2$ rather than 3), so the alternative sampling distribution overlaps the non-rejection region more and the test is less likely to reject H_0.

§ 8.4 Paired data

1. **(1)** $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} < 0$.

   (2) Let $X_{\text{diff},i}$ be the difference (before − after) for participant i. Under H_0,

   $$T = \frac{\overline{X}_{\text{diff}} - 0}{S_{\text{diff}}/\sqrt{n}} \sim t(9).$$

   (3) $\dfrac{s_{\text{diff}}}{\sqrt{n}} = \dfrac{2.9740}{\sqrt{10}} \approx 0.9404$, so $t = \dfrac{-1.8000}{0.9404} \approx -1.9140$.

   (4) p-value $= \mathbb{P}(T < -1.9140) \approx 0.0439$ with $T \sim t(9)$. Since $0.0439 < 0.05$, we reject H_0; there is statistically significant evidence that the exercise program increases muscle mass.

2. **(1)** $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} > 0$.

   (2) With $n_{\text{diff}} = 9$ paired differences, under H_0,

   $$T = \frac{\overline{X}_{\text{diff}} - 0}{S_{\text{diff}}/\sqrt{n_{\text{diff}}}} \sim t(8).$$

   (3) From the output $t = 2.8$ on $df = 8$, so p-value $= \mathbb{P}(T > 2.8) \approx 0.0116$. Since $0.0116 < 0.05$, we reject H_0; the data provide strong evidence that the 8-week diet program reduces men's body fat on average.

3. **(1)** With $X_{\text{diff}} = X_A - X_B$ the paired difference (before − after), under H_0

   $$T = \frac{\overline{X}_{\text{diff}} - 0}{S_{\text{diff}}/\sqrt{n_{\text{diff}}}} = \frac{\overline{X}_{\text{diff}}}{S_{\text{diff}}/\sqrt{12}} \sim t(11).$$

   (2) The alternative $H_A : \mu_{\text{diff}} < 0$ is lower-tailed, so the rejection region is $t < -t_{0.05}(11) = -1.7959$.

   (3) The observed statistic is $t = -1.5$ on $df = 11$, with p-value $= \mathbb{P}(T < -1.5)$ for $T \sim t(11)$. Since $-1.5 > -1.7959$, t does not fall in the rejection region, so we fail to reject H_0. At $\alpha = 0.05$ there is insufficient evidence that grip strength is greater after taking the supplement than before.

   (4) ②. The p-value is the left-tail probability $\mathbb{P}(T < -1.5)$, i.e. the small area to the *left* of -1.5.

§8.5 Difference of two means: unequal variances

1. **(1)**

   $$H_0 : \mu_A - \mu_B = 0, \qquad H_A : \mu_A - \mu_B \neq 0$$

   (2) Both population variances are known and $\overline{X}_A - \overline{X}_B \sim N(\mu_A - \mu_B, \sigma_A^2/n_A + \sigma_B^2/n_B)$, so under H_0,

   $$Z = \frac{\overline{X}_A - \overline{X}_B - 0}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{4}} \sim N(0, 1).$$

   (3) This is a two-sided test with $\alpha = 0.05$, so the rejection region is

   $$|z| > z_{0.025} = 1.96.$$

   (4) Since $\overline{x}_A = 455$ and $\overline{x}_B = 449.5455$,

   $$z = \frac{455 - 449.5455}{\sqrt{\frac{18}{9} + \frac{22}{11}}} = 2.7273.$$

   (5) The observed statistic $z = 2.7273$ lies in the rejection region $|z| > 1.96$, so we reject H_0.

   Conclusion: At significance level $\alpha = 0.05$, there is sufficient evidence to conclude that the average distance of the new battery differs from that of the existing battery.

2. **(1)** The claim is that Brand 1 has lower mean pH level, so the alternative is $\mu_1 - \mu_2 < 0$.

$$H_0 : \ \mu_1 - \mu_2 = 0, \qquad H_A : \ \mu_1 - \mu_2 < 0.$$

(2) Since the population standard deviations are known, a two-sample Z test is used.

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \ \frac{0.1^2}{50} + \frac{0.1^2}{50}\right).$$

Since $\frac{0.1^2}{50} + \frac{0.1^2}{50} = 0.0004$, under $H_0 : \mu_1 - \mu_2 = 0$,

$$\overline{X}_1 - \overline{X}_2 \sim N(0, \ 0.02^2).$$

(3) For a left-sided test at $\alpha = 0.05$, the critical value is $z_{0.05} = 1.645$. Reject H_0 if

$$\bar{x}_1 - \bar{x}_2 < -z_{0.05} \cdot 0.02 = -1.6449(0.02) = -0.0329.$$

Thus, the rejection region is $\bar{x}_1 - \bar{x}_2 < -0.0329$.

(4)

$$\bar{x}_1 - \bar{x}_2 = 3.4 - 3.5 = -0.1.$$

Since $-0.1 < -0.0329$, the observed difference lies in the rejection region. Therefore, we reject H_0.

Conclusion: At the significance level $\alpha = 0.05$, there is sufficient evidence to conclude that Brand 1 wine has a lower mean pH level than Brand 2 wine.

(5)

$$\begin{aligned}
\mathbb{P}(\text{Reject } H_0 \mid H_A \text{ is true}) &= \mathbb{P}(\overline{X}_1 - \overline{X}_2 < -0.0329 \mid \mu_1 - \mu_2 = -0.05) \\
&= \mathbb{P}\left(\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{0.02} < \frac{-0.0329 - (-0.05)}{0.02}\right) \\
&= \mathbb{P}(Z < 0.8550) \approx 0.8037.
\end{aligned}$$

3. **(1)** $H_0 : \mu_A - \mu_B = 0$ versus $H_A : \mu_A - \mu_B \neq 0$.

(2) Since the population variances are unknown and not assumed equal,

$$T = \frac{(\overline{X}_A - \overline{X}_B) - 0}{\sqrt{\dfrac{S_A^2}{n_A} + \dfrac{S_B^2}{n_B}}} \overset{H_0}{\sim} t(\psi),$$

where ψ is the Satterthwaite degrees of freedom.

(3) From the output, $t = -0.0669$ and p-value $= 0.9472$. Since $0.9472 > \alpha = 0.01$, we fail to reject H_0; there is insufficient evidence at the 1% level that the mean signal strength differs between antennas A and B.

§ 8.6 Difference of two means: equal variances

1. **(1)**

$$H_0 : \mu_A - \mu_B = 0, \qquad H_A : \mu_A - \mu_B \neq 0$$

(2) Under H_0, the test statistic is given by

$$T = \frac{\overline{X}_A - \overline{X}_B - 0}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{\overline{X}_A - \overline{X}_B}{S_p\sqrt{\frac{1}{6} + \frac{1}{6}}} \overset{H_0}{\sim} t(n_A + n_B - 2),$$

where $\overline{X}_A$ and $\overline{X}_B$ are the sample means of corn yields from fertilizers A and B, and $S_p^2 = \dfrac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$ is the pooled sample variance. Here $n_A + n_B - 2 = 10$.

(3) The observed test statistic is $t = 0.31311$ on $df = 10$, and the p-value is 0.7606.

Since the p-value is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

Conclusion: There is insufficient evidence to conclude a difference in mean corn yields between fertilizers A and B.

2. **(1)** $H_0 : \mu_E - \mu_R = 0$ versus $H_A : \mu_E - \mu_R \neq 0$.

(2) Assuming equal variances, under H_0,

$$T = \frac{\overline{X}_E - \overline{X}_R}{S_p\sqrt{\frac{1}{n_E} + \frac{1}{n_R}}} \overset{H_0}{\sim} t(df), \qquad df = n_E + n_R - 2 = 20.$$

(3) From the output $t \approx 3.3673$ on $df = 20$, so p-value $= 2\mathbb{P}(T > 3.3673) \approx 0.0030$. Since $0.0030 < 0.05$, we reject H_0; there is sufficient evidence that mean daily sugar intake differs between the two groups.

(4) Since $S_p^2 = \dfrac{(n_E - 1)S_E^2 + (n_R - 1)S_R^2}{n_E + n_R - 2}$, comparing the two expressions gives

$$V = \frac{(n_E + n_R - 2)S_p^2}{\sigma^2} \sim \chi^2(n_E + n_R - 2).$$

§8.7 Power calculations for a difference of means

1. **(1)** The claim is that Group 1 sleeps more on average, so the alternative is $\mu_1 - \mu_2 > 0$.

$$H_0 : \mu_1 - \mu_2 = 0, \qquad H_A : \mu_1 - \mu_2 > 0$$

(2) The sampling distributions are $\overline{X}_1 \sim N(\mu_1, 1/8)$, $\overline{X}_2 \sim N(\mu_2, 2.25/18)$. Thus,

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2,\ \frac{1}{8} + \frac{2.25}{18}\right).$$

Since $1/8 + 2.25/18 = 0.25$, under $H_0 : \mu_1 - \mu_2 = 0$,

$$\overline{X}_1 - \overline{X}_2 \sim N(0,\ 0.5^2).$$

(3) For a right-sided test at $\alpha = 0.05$, the critical value is $z_{0.05} = 1.645$. Reject H_0 if

$$\bar{x}_1 - \bar{x}_2 > z_{0.05} \cdot SE = (1.645)(0.5) = 0.8225.$$

Thus, the rejection region is $\bar{x}_1 - \bar{x}_2 > 0.8225$.

(4)

$$\bar{x}_1 - \bar{x}_2 = 8.2 - 7.0 = 1.2.$$

Since $1.2 > 0.8225$, the observed difference lies in the rejection region. Therefore, we reject H_0.

Conclusion: At the significance level $\alpha = 0.05$, there is sufficient evidence to conclude that people who exercise regularly sleep more (on average) than people who do not exercise regularly.

(5)

$$\begin{aligned}\mathbb{P}(\text{Reject } H_0 \mid H_A \text{ is true}) &= \mathbb{P}(\overline{X}_1 - \overline{X}_2 > 0.8225 \mid \mu_1 - \mu_2 = 1) \\ &= \mathbb{P}\left(\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{0.5} > \frac{0.8225 - 1}{0.5}\right) \\ &= \mathbb{P}\left(Z > -0.355\right) \approx 0.6387.\end{aligned}$$

2. **(1)** $H_0 : \mu_A - \mu_B = 2$ versus $H_A : \mu_A - \mu_B > 2$.

   (2) With known variances, $\overline{X}_A - \overline{X}_B \sim N\left(2, \frac{0.25}{40} + \frac{0.36}{50}\right)$ under H_0 (the standard error is $\sqrt{0.25/40 + 0.36/50} \approx 0.116$).

   (3) This is a right-tailed test with $z_{0.01} = 2.3263$, so the rejection region is

   $$\bar{x}_A - \bar{x}_B > 2 + 2.3263\sqrt{\tfrac{0.25}{40} + \tfrac{0.36}{50}} \approx 2.2698.$$

   (4) The observed value is $\bar{x}_A - \bar{x}_B = 27.48 - 25.17 = 2.31$, which lies in the rejection region, so we reject H_0. There is sufficient evidence at $\alpha = 0.01$ that the new earplugs reduce noise by more than 2 dB on average.

   (5) The power is

   $$1 - \beta = \mathbb{P}\left(\overline{X}_A - \overline{X}_B > 2.2698 \,\middle|\, \mu_A - \mu_B = 2.5\right) = \mathbb{P}\left(Z > \frac{2.2698 - 2.5}{0.116}\right) = \mathbb{P}(Z > -1.9850) \approx 0.9764.$$

§8.8 Comparing many means: ANOVA

1. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_A : not all four lecture means are equal. With $k = 4$ groups and $N = 4 \times 50 = 200$ observations, $df_G = k - 1 = 3$ and $df_E = N - k = 200 - 4 = 196$.

2. $$\text{MSG} = \frac{\text{SSG}}{df_G} = \frac{1290}{2} = 645, \quad \text{MSE} = \frac{\text{SSE}}{df_E} = \frac{29810}{161} \approx 185.16, \quad F = \frac{\text{MSG}}{\text{MSE}} = \frac{645}{185.16} \approx 3.48.$$

3. There are $\binom{4}{2} = 6$ pairwise comparisons, and the Bonferroni level is $\alpha^\star = \dfrac{\alpha}{6} = \dfrac{0.05}{6} \approx 0.0083$.

4. **(1)** $H_0 : \mu_1 = \mu_2 = \cdots = \mu_8$ versus H_A : not all the μ_i are equal.

   (2) Degrees of freedom: Company $= k - 1 = 8 - 1 = 7$; Residuals $= n - k = 268 - 8 = 260$. Then

   $$\text{MSG} = \frac{12.9853}{7} \approx 1.8550, \quad \text{MSE} = \frac{17.5223}{260} \approx 0.0674, \quad F = \frac{1.8550}{0.0674} \approx 27.5223.$$

(3) Since the p-value (< 0.001) is less than $\alpha = 0.05$, we reject H_0. There is strong statistical evidence that cadet life satisfaction differs across the eight companies.

5. **(1)** Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the true mean GPA for the Humanities, Natural Sciences, Social Sciences, and Engineering groups. Then $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_A : not all the μ_j are equal.

   (2) With $k = 4$ and $N = 268$: Major $df = k - 1 = 3$ and MSG $= \dfrac{0.0348}{3} = 0.0116$; Residuals $df = N - k = 264$ and MSE $= \dfrac{20.6831}{264} = 0.0783$; and $F = \dfrac{0.0116}{0.0783} \approx 0.1481$. (Filled cells, left to right: 3, 0.0116, 0.1481, 264, 0.0783.)

   (3) The p-value is $0.931 > \alpha = 0.05$, so we fail to reject H_0. There is no evidence that the average GPA differs across the four majors.

Chapter 9

§9.1 Line fitting, residuals, and correlation

1. **(1)**
$$H_0 : \rho = 0, \qquad H_A : \rho > 0$$

   (2) With $n = 31$, under H_0,
$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{R\sqrt{29}}{\sqrt{1-R^2}} \sim t(29).$$

   (3) This is a right-sided test, so the rejection region is $t > t_\alpha(n-2)$. With $\alpha = 0.05$ and $n = 31$,
$$t > t_{0.05}(29) = 1.6991.$$

   (4) The observed test statistic is $t = 4.0776$ with $df = n - 2 = 29$. For a right-sided test,
$$p\text{-value} = \mathbb{P}(T > 4.0776) = 0.0002.$$
   Since the p-value is smaller than $\alpha = 0.05$, we reject H_0.

   Conclusion: At significance level $\alpha = 0.05$, there is sufficient evidence of a positive correlation between the summer daily maximum temperature and the café's delivery sales.

2. **(1)** $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$.

   (2) Under H_0, with $n = 42$,
$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(40).$$

(3) $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{319.2}{\sqrt{361 \cdot 576}} = \dfrac{319.2}{456} = 0.70$, so
$$t = \frac{0.70\sqrt{40}}{\sqrt{1-0.70^2}} \approx 6.1993.$$

(4) p-value $= 2\mathbb{P}(T > 6.1993) \approx 2.48 \times 10^{-7}$. Since this is far smaller than $\alpha = 0.05$, we reject H_0; there is strong evidence of a (positive) correlation between age and annual medical charges.

3. **(1)** $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$.

   (2) With $n = 10$, under H_0,

   $$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{R\sqrt{8}}{\sqrt{1-R^2}} \sim t(8).$$

   (3) This is a two-sided test, so the rejection region is $|t| > t_{0.025}(8) = 2.3060$.

   (4) ③. Since $r = 0.9652$ is close to 1, the points show a strong positive linear pattern.

   (5) The observed test statistic is

   $$t = \frac{r\sqrt{8}}{\sqrt{1-r^2}} = \frac{0.9652\sqrt{8}}{\sqrt{1-0.9652^2}} \approx 10.4393.$$

   Since $|t| = 10.4393 > 2.3060$, we reject H_0. At $\alpha = 0.05$ there is sufficient evidence of a correlation between the floor area and the weekly sales.

4. **(1)** $H_0 : \rho = 0$ versus $H_A : \rho > 0$.

   (2) With $n = 11$, under H_0,

   $$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{R\sqrt{9}}{\sqrt{1-R^2}} \sim t(9).$$

   (3) $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \dfrac{26.6}{\sqrt{418 \cdot 2.3418}} \approx 0.8502.$

   (4) $\dfrac{r\sqrt{11-2}}{\sqrt{1-r^2}} = \dfrac{0.8502 \cdot 3}{\sqrt{1-0.8502^2}} \approx 4.8448.$

   (5) p-value $= \mathbb{P}(T > 4.8448) \approx 0.0004$ with $T \sim t(9)$. Since the p-value is less than $\alpha = 0.05$, we reject H_0. There is statistically significant evidence of a positive linear relationship between game time and salary.

§ 9.2 Fitting a line by least squares

1. $\hat{\beta}_1 = R\dfrac{s_y}{s_x} = (-0.6)\dfrac{4}{10} = -0.24$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 120 - (-0.24)(50) = 132$. The fitted line is $\hat{y} = 132 - 0.24x$.

2. $R^2 = 0.8^2 = 0.64$, so the model explains 64% of the variation in y. The remaining 36% $(1 - R^2)$ is the variation in y that the linear fit leaves unexplained (the residual/error variation).

3. Setting `is_premium` $= 0$ gives the non-premium group mean $\hat{y} = 30$; setting `is_premium` $= 1$ gives the premium group mean $\hat{y} = 30 + 7 = 37$.

4. **(1)** $\hat{\beta}_1 = r\dfrac{s_y}{s_x} = (-0.80)\dfrac{3000}{2} = -1200$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 14000 - (-1200)(6) = 21200$.

   (2) $\hat{y}_i = 21200 - 1200(9) = 10400$ dollars, so $e_i = y_i - \hat{y}_i = 10000 - 10400 = -400$ dollars.

   (3) $R^2 = r^2 = (-0.80)^2 = 0.64$. With $\text{SST} = (n-1)s_y^2 = 31(3000^2) = 279{,}000{,}000$ and $\text{SSE} = \text{SST}(1 - R^2) = 100{,}440{,}000$,

   $$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{100{,}440{,}000}{30}} \approx 1829.75.$$

5. **(1)** $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{4.414}{45.45} \approx 0.0971$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 3.8 - 0.0971(49.825) \approx -1.0380$.

   (2) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(50) = -1.0380 + 0.0971(50) \approx 3.8170$ kg.

   (3) Since $x_i' = 10x_i$, we have $S_{x'x'} = 100\,S_{xx} = 4545$ and $S_{x'y} = 10\,S_{xy} = 44.14$, so

$$\hat{\beta}_1' = \frac{S_{x'y}}{S_{x'x'}} = \frac{44.14}{4545} \approx 0.0097, \qquad \hat{\beta}_0' = \overline{y} - \hat{\beta}_1'\overline{x}' = 3.8 - 0.0097(498.25) \approx -1.0380.$$

   The intercept is unchanged, and the slope is scaled by 1/10 (consistent with the change of units).

6. **(1)** $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{26.6}{418} \approx 0.0636$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 2.6727 - 0.0636(20) \approx 1.4007$.

   (2) $\hat{y}_4 = 1.4007 + 0.0636(25) \approx 2.9907$, so $e_4 = y_4 - \hat{y}_4 = 3.0 - 2.9907 = 0.0093$.

   (3) $MSE = \dfrac{SSE}{n-2} = \dfrac{0.6489}{11-2} \approx 0.0721$.

7. **(1)** $\hat{\beta}_1 = r\dfrac{s_y}{s_x} = 0.64\dfrac{60}{480} = 0.08$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 210 - 0.08(1400) = 98$.

   (2) $\hat{y}_i = 98 + 0.08(1000) = 178$ minutes, so $e_i = y_i - \hat{y}_i = 190 - 178 = 12$ minutes.

   (3) $R^2 = r^2 = 0.64^2 = 0.4096$. With $\text{SST} = (n-1)s_y^2 = 29(60^2) = 104{,}400$,

$$\text{SSE} = \text{SST}(1 - R^2) = 104{,}400(1 - 0.4096) = 61{,}637.76.$$

§9.3 Inference for linear regression

1. **(1)** $\hat{\beta}_0 = 52.5152$, $\hat{\beta}_1 = 1.3939$, and the estimated regression equation is $\hat{Y} = 52.5152 + 1.3939\,x$.

   (2) By the estimated regression equation, the fitted value is

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 52.5152 + 1.3939 \times 10 \approx 66.4542,$$

   and the residual is

$$e_1 = y_1 - \hat{y}_1 = 64 - 66.4542 = -2.4542.$$

   (3)

$$\sum_{i=1}^{10} e_i^2 = SSE = (10-2) \times MSE = 8 \times 102.01 = 816.08.$$

   (4) a. test statistic = 2.506, p-value = 0.03659; therefore we **reject** H_0 at $\alpha = 0.05$.

   b. At significance level $\alpha = 0.05$, the explanatory variable (practice time) is a statistically significant predictor of the response variable (shooting-test score).

2. Since

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad E(\varepsilon_i) = 0,$$

   we have

$$E(Y_i) = \beta_0 + \beta_1 x_i.$$

Taking expectation of $\hat{\beta}_1$,

$$E(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \sum_{i=1}^{n}(x_i - \overline{x})E(Y_i)$$
$$= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i).$$

Using $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$,

$$\sum_{i=1}^{n}(x_i - \overline{x})\beta_0 = 0,$$

so

$$E(\hat{\beta}_1) = \frac{\beta_1 \sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Since

$$\sum_{i=1}^{n}(x_i - \overline{x})x_i = \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

it follows that

$$E(\hat{\beta}_1) = \beta_1.$$

Therefore, $\hat{\beta}_1$ is an unbiased estimator of β_1.

3. **(1)** From the output $\hat{\beta}_0 = 39.1$ and $\hat{\beta}_1 = -6.3$. For smokers ($x = 1$), $\hat{y} = 39.1 - 6.3 = 32.8$; for non-smokers ($x = 0$), $\hat{y} = 39.1$.

   (2) $\overline{x} = \dfrac{5(1) + 10(0)}{15} = \dfrac{5}{15} \approx 0.3333$, and using $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$, $\overline{y} = 39.1 + (-6.3)(0.3333) \approx 37.0002$.

   (3) Since $R^2 = 0.294$ and $\hat{\beta}_1 = r\dfrac{s_y}{s_x} < 0$, the correlation is negative: $r = -\sqrt{0.294} \approx -0.5422$.

   (4) (i) $T = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(13)$, since $df = 15 - 2 = 13$. (ii) The observed statistic is -2.327; the two-sided p-value is 0.0368, so the one-sided p-value is $\mathbb{P}(T < -2.327) = 0.0184$. Since $0.0184 < 0.05$, we reject H_0.

§ 10.1 Modeling a binary outcome

1. **(1)** For $p = 0.8$, the odds are $\dfrac{0.8}{1 - 0.8} = \dfrac{0.8}{0.2} = 4$, and the logit is $\log(4) \approx 1.386$.

   (2) If the odds are $1/4$, then $p = \dfrac{\text{odds}}{1 + \text{odds}} = \dfrac{1/4}{1 + 1/4} = \dfrac{1/4}{5/4} = \dfrac{1}{5} = 0.2$.

   (3) A logit of 0 gives $p = \sigma(0) = \dfrac{1}{1 + e^0} = 0.5$. A logit of -2 gives $p = \sigma(-2) = \dfrac{1}{1 + e^2} \approx 0.119$.

2. **(1)** At $x = 5$, $\text{logit}(\hat{p}) = -1.5 + 0.5(5) = 1.0$, so $\hat{p} = \sigma(1.0) = \dfrac{1}{1 + e^{-1.0}} \approx 0.731$.

   (2) $\hat{p} = 0.5$ exactly when the logit is 0: $-1.5 + 0.5x = 0$, so $x = 3$.

§ 10.2 Fitting and interpreting logistic regression

1. **(1)** The income coefficient is negative (-0.04), so higher income is associated with smaller log-odds — and hence a smaller probability — of default.

   (2) Income is measured in thousands of dollars, so a \$10,000 increase is 10 units. The odds multiply by $e^{-0.04\times 10} = e^{-0.4} \approx 0.670$: each additional \$10,000 of income multiplies the odds of default by about 0.67, a 33% reduction in the odds.

   (3) For income = 50, $\hat{z} = 2.0 - 0.04(50) = 0$, so $\hat{p} = \sigma(0) = 0.5$.

2. **(1)** $\hat{z} = -4.00 + 0.0030(1200) = -0.4$, so $\hat{p} = \sigma(-0.4) = \dfrac{1}{1+e^{0.4}} \approx 0.401$.

   (2) Each additional \$100 raises the log-odds by $0.0030 \times 100 = 0.30$, so the odds multiply by $e^{0.30} \approx 1.35$ — a 35% increase in the odds of fraud.

§ 10.3 Making and evaluating decisions

1. **(1)** Sensitivity $= \dfrac{TP}{TP+FN} = \dfrac{80}{80+20} = 0.80$; specificity $= \dfrac{TN}{TN+FP} = \dfrac{360}{360+40} = 0.90$.

   (2) Accuracy $= \dfrac{TP+TN}{n} = \dfrac{80+360}{500} = \dfrac{440}{500} = 0.88$.

   (3) Precision $= \dfrac{TP}{TP+FP} = \dfrac{80}{80+40} = \dfrac{80}{120} \approx 0.667$.

2. **(1)** Of 10,000 people, 2% = 200 have the disease and 9,800 do not. The test produces $TP = 0.95 \times 200 = 190$ true positives and $FP = (1-0.90) \times 9{,}800 = 0.10 \times 9{,}800 = 980$ false positives.

   (2) $\mathbb{P}(\text{disease} \mid \text{positive}) = \dfrac{TP}{TP+FP} = \dfrac{190}{190+980} = \dfrac{190}{1170} \approx 0.162$. Even with high sensitivity and specificity, only about 16% of positive tests are correct, because the disease is rare: the 10% false-positive rate applied to the large healthy majority swamps the true positives.

3. **(1)** Model A (AUC 0.80) ranks a random positive above a random negative more reliably than model B (AUC 0.55).

   (2) An AUC of 0.5 corresponds to random guessing (chance-level ranking). An AUC below 0.5 indicates systematically reversed ranking — the model tends to score negatives higher than positives, so flipping its scores would do better than chance.

§ 10.4 From regression to machine learning

1. The model is fit to make the training outcomes as probable as possible (equivalently, to make the training error small), so it adapts to the particular noise and quirks of the training sample. Part of that fit reflects patterns that will not recur in new data, so training accuracy is optimistically biased. Accuracy on a held-out test set — data not used in fitting — is the honest estimate of performance on new cases.

2. **(1)** The flexible model is overfitting: near-perfect training accuracy (99%) but much lower test accuracy (72%). The large gap between training and test performance is the signature of high variance (memorizing noise rather than signal).

   (2) The simpler model should generalize better. Its test accuracy (83%) is higher and close to its training accuracy (85%), showing it captures real structure without overfitting.

3. The bias–variance tradeoff is that more flexible models can represent more complex patterns (lowering bias) but are more sensitive to the particular training sample (raising variance), while simpler models do the reverse; the best generalization balances the two. Increasing model flexibility tends to *decrease* the bias and *increase* the variance.

References

Agresti, A., Franklin, C., & Klingenberg, B. (2022). *Statistics: The Art and Science of Learning from Data* (5th ed.). Pearson.

Casella, G., & Berger, R. (2024). *Statistical Inference*. Chapman and Hall/CRC.

Çetinkaya-Rundel, M., & Hardin, J. (2024). *Introduction to Modern Statistics* (2nd ed.). OpenIntro.

Diez, D. M., Barr, C. D., & Çetinkaya-Rundel, M. (2019). *OpenIntro Statistics* (4th ed.). OpenIntro.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). W. W. Norton.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Ismay, C., Kim, A. Y., & Valdivia, A. (2025). *Statistical Inference via Data Science: A ModernDive into R and the Tidyverse* (2nd ed.). Chapman and Hall/CRC.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.

Kang, J., Bang, S., Shin, J., & Lee, Y. (2022). *Understanding and Application of Statistics with R (Korean)*. Seoul, Korea: Kyowoo.

Kim, J. K. (2026). *Statistics in Survey Sampling*. CRC Press.

Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.

Ross, S. M. (1998). *A First Course in Probability* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Ross, S. M. (2020). *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical Methods* (8th ed.). Ames, Iowa: Iowa State University Press.

Index